Dynamic Polycomb Chromatin Suppresses Aberrant Transcription in *Drosophila* Immunity

DISSERTATION
zur Erlangung des akademischen Grades

Doctor rerum naturalium
(Dr. rer. nat.)

eingereicht an der
Lebenswissenschaftlichen Fakultät der Humboldt-Universität zu Berlin

von
M.Sc. Robert Streeck

Präsidentin/Präsident
der Humboldt-Universität zu Berlin

Prof. Dr.-Ing. Dr. Sabine Kunst

Dekanin/Dekan der Lebenswissenschaftlichen Fakultät
der Humboldt-Universität zu Berlin

Prof. Dr. Bernhard Grimm

Gutachter/innen

1. Prof. Arturo Zychlinsky, PhD
2. Prof. Dr. Leonie Ringrose
3. Prof. Dr. Daniel Schubert

Tag der mündlichen Prüfung:
29.06.2020

1 TABLE OF CONTENTS

2 ZUSAMMENFASSUNG

Chromatin ist mehr als ein inerter Komplex, der genetische Information speichert. Die Modifikation von Chromatin und Histonen sind vielmehr zentrale Ereignisse in der Regulation der Geneexpression. Es ist gut etabliert, dass die Trimethylierung von Histon H3 Lysin 27 (H3K27me3) durch ‚Polycomb group' (PcG) Proteine den reprimierten Zustand von entwicklungsbiologisch relevanten Transkriptionsfaktoren als epigenetisches Gedächtnis aufrechterhält. Genomweite Analysen haben jedoch gezeigt, dass ein großer Teil des nicht-exprimierten Genoms die H3K27me3 Modifikation trägt. Daraus ergibt sich die Frage ob diese Histonmodifikation auch außerhalb des Entwichlungs-Zusammenhangs funktional relevant ist um die Aktivierung von Genen durch physiologische Signalwege zu verhindern.

In dieser Arbeit habe ich RNA-seq und ChIP-seq auf Plasmatozyten, Makrophagen-artige Immunzellen aus *Drosophila*, angewendet und ein hochauflösendes Chromatinprofil sowie eine transkriptionelle Blaupause dieser homogenen Zellpopulation generiert. Dadurch konnte ich zeigen, dass eine Gruppe von Immungenen, die durch H3K27me3 markiert sind, nach einer Immunstimulierung rasch hochreguliert wird. Weiterhin konnte ich durch die Anwendung eines neu entwickelten Genomanalyse-Algorithmus zeigen, dass diese H3K27me3 positiven Gene in einem Chromatinzustand sind, der sich von kanonischem Polycomb Chromatin unterscheidet. Dieser Chromatinzustand hat zwei wichtige Eigenschaften: Erstens, indem ich zeige, dass H3K27me3 in Plasmatozyten nach einer Immunaktivierung an Immungenen verloren geht, beweise ich, dass H3K27me3 als Antwort auf physiologische Stimuli dynamisch reguliert wird. Zweitens, indem ich etabliere, dass Immungene hochreguliert werden, wenn H3K27me3 durch Mutationen in PcG Proteinen oder Histon H3 reduziert wird, weise ich nach, dass es für den Erhalt eines reprimierten Genezustands instruktiv ist. Daher bezeichne ich diesen Chromatinzustand als dynamisches Polycomb Chromatin.

Meine weiteren Analysen haben gezeigt, dass dieses dynamisches Polycomb Chromatin in *Drosophila* Plasmatozyten mit einer großen Zahl anderer dynamisch regulierter Gene assoziieret ist. Einige dieser Gene wurden ebenfalls nach der Depletion von H3K27me3 durch genetische Manipulation hochreguliert. Deshalb schlage Ich vor, dass dynamisches Polycomb Chromatin einen neuartigen Chromatinzustand darstellt, der an Genen zu finden ist, deren Transkription durch unvorhersehbare Ereignisse ausgelöst wird. Die Reprimierung durch dynamisches Polycomb Chromatin bildet demnach eine Aktivierungschwelle gegen aberrante Transkription, die aber trotzdem eine rasche Geninduktion in physiologisch relevanten Situationen erlaubt.

3 ABSTRACT

Chromatin is more than an inert complex that stores genetic information. Modification of chromatin and histones are central events in regulating gene expression. It is clearly established that histone H3 lysine 27 trimethylation (H3K27me3) by Polycomb group (PcG) proteins maintains the repressed state of developmental transcription factors in epigenetic memory. Genome wide analysis revealed, however, that much of the non-transcribed genome carries the H3K27me3 modification. This raises the question, whether this modification is functionally relevant outside of development to preclude activation of genes through physiological signaling, analogous to its function in epigenetic memory.

Here, I applied RNA-seq and ChIP-seq to *Drosophila* plasmatocytes, a macrophage-like immune cell, generating a high resolution chromatin landscape and transcriptional blueprint of this homogeneous cell population. Thereby, I demonstrated that a set of H3K27me3 marked immune genes was rapidly transcriptionally induced upon challenge. Further, by applying a newly developed genome clustering algorithm, I demonstrated that these H3K27me3 positive genes are in a chromatin state that is distinct from the canonical Polycomb chromatin, in which developmental transcription factors are silenced. This state has two important properties: First, by demonstrating that in plasmatocytes H3K27me3 is depleted specifically at immune genes after activation, I showed that it is dynamically regulated in response to physiological stimulation. Second, by establishing that immune genes were up-regulated when H3K27me3 was depleted by mutating PcG proteins or histone H3, I confirmed that it is instructive in maintaining a silenced gene state. Therefore, I termed this novel chromatin state dynamic Polycomb chromatin.

Further analysis revealed that this dynamic Polycomb chromatin state is also associated with a large number of other dynamically regulated genes in *Drosophila* plasmatocytes. Some of these genes were also up-regulated when H3K27me3 is depleted by genetic manipulation. Hence, I propose that dynamic Polycomb chromatin is a novel chromatin state that targets genes which, in contrast to developmental genes, are triggered by non-predetermined signaling events. Silencing by dynamic Polycomb chromatin thresholds such genes against aberrant gene activation, but permits rapid induction in physiologically relevant situations.

4 INTRODUCTION

4.1 Histone modifications and transcriptional regulation

4.1.1 TRANSCRIPTION AND CHROMATIN

Our genes carry the blueprint that make up the cells in our body. The main actors, that define the function of these cells are however RNAs and proteins. To produce these our genes are transcribed into mRNAs by RNA polymerases and translated to proteins by ribosomes. But if our genes are the same across all cells of our body, how can these cells be so different and carry out so many different functions?[1-3] Therefore, factors must exist that modulate the transcription (and translation) of genes to achieve the appropriate activity of a gene at any given time and in any given cell.

It is well established that transcription factors are one major mechanism by which transcription can be modulated in a temporal and context dependent manner. When activated these transcription factor proteins recognize specific DNA sequences which they bind triggering transcription of adjacent genes.[4] This basic mechanism of gene control is well conserved across prokaryotes and eukaryotes and helps cells maintain cell states and respond to environmental changes.[5,6]

Another crucial player in gene regulation is chromatin. It is a complex of DNA and a large number of proteins found in the nucleus of eukaryotic cells.[7] The smallest unit in chromatin is the nucleosome. It organizes 146 base pairs of DNA, which are wrapped in 1.67 superhelical turns around an octamer of highly conserved histone proteins – containing 2 copies of each of the core histones H2A, H2B, H3 and H4.[8] This basic unit is then organized into higher structures by other proteins, including the linker histone H1, that together make up chromatin.[9] This structure aids in the 3D organization and storage of the large eukaryotic genomes.

Chromatin is however much more than an inert platform that packages DNA. It is in fact heavily modified and varies greatly in its composition, density and modifications both across genomic regions and between individual cells. [10-13]

4.1.2 HISTONE MODIFICATIONS

Histones in particular undergo a tremendous number of posttranslational modifications (PTMs). In this regard, the N-terminal tails of histones are of particular interest. While the globular core of the histone octamer is tightly wrapped by DNA, the long and unstructured N-terminal parts of the histone subunits protrude outwards and are accessible for protein-protein interactions and posttranslational modification. [14,15]

Among the most studied histone PTMs are histone acetylation and histone methylation[16,17], but many further histone modifications like ubiquitination, phosphorylation and citrullination are also well documented.[14,15] In fact, histone H3 alone has 5 lysines that carry well studied methylations and various approaches have been used to identify many more histone lysine and arginine residues that carry posttranslational modifications. [18,19]

Most of these histone lysine methylations can be either mono-, di- or trimethylations.[18,19] These 'marks' are placed by one or more methyltransferases – so called histone writers - that are often highly specific for the residue that they modify. [20-24] The writers may either catalyze one step in the process of histone modification – only adding one methyl group towards mono-, di- or trimethylation[21,22] - or perform several steps at once[23]. Histone methylation is very stable and is therefore considered critical in maintaining information across longer time spans and through cell divisions.[25]

Histone lysine acetylation also occurs on a large number of residues.[17] Here only a single acetyl group is being transferred by its writer enzymes.[26-28] Histone acetylation is thought to be much shorter lived than histone methylation, but it is of high interest because it changes the charge properties of the histone: By neutralizing the positive charge of the lysine being modified it weakens the affinity of the normally positively charged histone to the negatively charged DNA. Because the now less positively charged N-terminus can no longer bind adjacent DNA as efficiently, this change in affinity could then result in an overall decondensation of chromatin, allowing better access for other protein complexes that translate or replicate the DNA bound by the modified nucleosome.[29-31] There exist however also dedicated proteins that bind to existing histone acetylations (readers), and it remains unclear, to what extent either the charge property or the or the reader binding confers the function of individual acetylations.[32,33]

4.1.3 THE HISTONE CODE

The advent of new genome-wide profiling methods has led to a greatly improved understanding of the localization of histone modifications and other proteins across the genome. The use of Next generation sequencing methods in particular has allowed the localization histone modifications with almost single nucleosome precision.[34-36] This has led to the observation that histone marks correlate with the transcriptional activity of associated genes and are distinctly located on specific genetic elements – enhancers, promotors or exons for example [37,38] Here I will focus on the proteins and modifications as they are described in the model system used in this thesis - *Drosophila melanogaster*. These processes are however well conserved across metazoans and orthologues of all enzymes can be found in mammals.[39]

The main focus in this thesis lies on Histone H3 Lysine 27 trimethylation (H3K27me3) and the related Histone H3 Lysine 27 dimethylation (H3K27me2). In *Drosophila*, both marks are processed by the enzyme Enhancer of Zeste (E(z))[40,41] which, as part of the Polycomb repressive Complex 2 (PRC2) places these modifications at and around non-transcribed genes. H3K27 methylation is therefore found both along the whole gene body – including introns and exons – but also in intergenic regions around silenced genes.[42-44] Because of the relevance of this process in this thesis I will return to H3K27me3 and Polycomb later in their own chapter.

Another mark that is frequently associated with silent (non-transcribed) chromatin but also with structural chromatin (e.g. telomeres and centromeres) is Histone H3 Lysine 9 trimethylation.[45] This mark is placed, with the help of further proteins and complexes, by the enzyme Suppressor of Variegation 3-9 (Su(var)3-9)[46] and is a hallmark of constitutive heterochromatin – a type of highly compacted chromatin.[47] This heterochromatin is bound by specific proteins like Heterochromatin Protein 1 [48-50] and is therefore either silent or structural.[51]

Further, I interrogated a number of modifications that are frequently associated with actively transcribed genes. One of these is Histone H3 Lysine 9 acetylation (H3K9ac), which is placed by the Gcn5 Acetyltransferase as part of the SAGA complex.[52,53] It is found both at active enhancers and around the transcriptional start sites (TSS) of active genes.[54,55] Similarly distributed is the other histone acetylation I investigated in this thesis - Histone H3 Lysine 27 acetylation (H3K27ac).[55] It is placed by the acetyltransferase Nejire (nej, also called dCBP).[28]

Additionally, I investigated a number of methylations that are associated with actively transcribed genes. Histone H4 lysine 20 monomethylation (H4K20me1) is placed by PR/SET Domain Containing Protein 7 [56,57] and is found at promotors and along the gene body of active genes.[58,59]

Next there is a set of methylations on the tail of histone H3 which are associated with active transcription. I tested Histone H3 lysine 4 for both mono- and trimethylation (H3K4me1 and H3K4me3), which are both placed by distinct enzymes. While H3K4me1 is placed by Trithorax-Related[22] or Trithorax[60] and is found at enhancers, predominantly those of active genes[55], H3K4me3 is placed by SET Domain Containing 1 [21,61] and is found around transcriptional start sites (TSS) of active genes. [54,55] I also assayed histone H3 lysine 36 trimethylation (H3K36me3), which is written by Absent, Small, or Homeotic Discs 1 (ash1) [62-64] and is found on exons of active genes, particularly towards the 3' end of the gene.[57,65]

Among these last writers, the enzymes that write H3K27ac, H3K4me1, H3K4me3 and H3K36me3 are all part of a group of proteins that are called Trithorax Group proteins (TrxG)[66] which have

been well described to be involved in gene activation, specifically during development where they counteract Polycomb based repression. [66]

This large diversity of histone modifications – I limited myself to the ones reported in this thesis – with their specificity towards certain genes and their characteristic localization on certain functional genomic elements has led to the formulation of the histone code hypothesis. [67,68] It proposes, that the histone PTMs can fill the gap in long term transfer of transcriptional information and encode cell specific information that can – unlike DNA – not only be read but also dynamically written, thereby "considerably extending the information potential of the genetic (DNA) code".[67]

Therefore, considerable effort has been made to integrate a multitude of histone PTMs as well as additional chromatin bound proteins into a single model, proposing the existence of distinct chromatin states defined by the combination of histone modifications present on them. [55,69] In Kharchenko, et al. [55] they split the genome in 200-bp windows and used a Hidden Markov Model (HMM) to fit 9 different chromatin types, based on a number of ChIP-seq experiments performed in *Drosophila* cell lines. By doing this, they were able to identify chromatin types corresponding to genetic elements like enhancers, TSS or active gene bodies with characteristic enrichment of histone modifications along those elements. In Filion, et al. [69] they instead used DamID [70] to identify protein binding along genomic regions again using an HMM model to fit 5 chromatin states. These states, being based entirely on protein binding data, reflected the gene state and also associated with the characteristic histone modifications of that state. Both Kharchenko, et al. [55] and Filion, et al. [69] acknowledge however, that the number of states in their studies were chosen arbitrarily and were selected such that they best corresponded to their prior knowledge of chromatin structure and function.

These types of studies come with a caveat however. While the presence of repressive histone modifications may strongly correlate with absence of transcription and the presence of active histone marks may be sufficient to predict of active transctiption[54,55,71], these observations still only demonstrate a correlation of transcription and histone modifications. Therefore, they cannot inform about the order of events, the instructiveness of each modification and whether or not they might be required to maintain an active or repressed gene state or just follow as a consequence of that gene state.[71,72]

4.1.4 Mutating histone posttranslational modifications

In order to address the instructive role of histone posttranslational modifications in transcription it was necessary to manipulate these PTMs – for example preventing them to be placed – and

identify changes in transcription as a consequence of this manipulation. One approach to prevent histone marks from being placed is mutating or inhibiting the enzyme or enzymes responsible for writing them.[21,40,60-62] Ablating histone PTMs in this fashion is achievable in a wide number of organism and cellular settings, it does however have a few shortcomings: The enzymes in question may also modify different targets[73-75] or fulfil functions other than the enzymatic processing of posttranslational modifications, whether alone or as part of a protein complex. [60,76-79]

Therefore, researchers have sought to also mutate the modified histone residues. This is faced with some challenges however. While in *Saccharomyces cerevisiae* it is possible to generate histone mutants[80,81], this did not appear to vastly alter transcriptional patterns, and the question remained, whether this would be different in a more complex, multicellular metazoan, where generating such histone mutants has proved much more difficult. While in metazoans the genes encoding for the histones H2A, H2B, H3 and H4 are commonly found in clusters, these clusters are spread across several loci throughout the genome.[82,83] In humans for example they are located on 3 different clusters across different chromosomes [82,84,85], which makes genetic manipulation of them difficult, if not impossible.

The canonical histones in *Drosophila* are however arranged in a distinct fashion: They are clustered in a single locus on chromosome 2 termed the histone complex, which is a tandem repeat of histone gene repeat units that in turn contain the 4 canonical core histones H2A, H2B, H3 and H4 as well as the linker histone H1.[86] This made it possible to generate mutant animals and cellular mosaics deficient for canonical histones which can then be rescued by insertion of transgenic histone gene units.[84,85] By using transgenes carrying point mutations in histones researchers used this *Drosophila* model to study the functional relevance of many histone PTMs, including H3K9 modifications[87], H3K36 modifications[65], H3K4 modifications[88] and H3K27 modifications.[25,89]

When looking at the results of these histone point-mutations it becomes clear however, that many seem to not show the expected changes in the transcriptional output that might be expected by their genomic distribution or by mutations in their writers – a derepression of silenced genes or a silencing of transcribed genes.[65,87,88] Even more, the histone mutants[87,88,90] often do not recapitulate the lethality phenotype of the writer enzyme mutants.[21,91,92] This might just be an effect of maternal contribution – the histone mRNAs or proteins present in the embryo deriving from the heterozygous mother – supplying sufficient histones to continue past critical developmental stages[93], but it might also demonstrate the further function of those write enzymes beyond the enzymatic function.

In conclusion, the histone replacement model in *Drosophila*, just like the mutations in histone writer enzymes, has its shortcomings. For each modification, one has to be wary of alternative modification on the same residue, for example lysine methylation and acetylation. Therefore, the best approach to identify the relevance of any one histone modification must be to determine the effects of both mutation in histone writers and histones themselves and integrate these findings to build a concordant model.

4.1.5 Polycomb Group Complexes And Histone H3 Lysine 27 trimethylation

One instance where the phenotypic changes and loss of repression in histone point-mutant and writer enzyme mutant correlated well is the H3K27me3/E(z) pair.[25,89] Because of its relevance in this thesis I want to go into more detail about how these act, with a particular focus on my model system *Drosophila*.

Much work on the mechanism of their gene silencing properties has focused on homeotic transformation, the full or partial transformation of one body part into another, and the misexpression of the homeobox transcription factors, particularly the subset of Hox transcription factors, which are critical for developmental patterning.[66,94] Hox genes are a set of transcription factors which maintain the segmental identity along the anterior-posterior axis such that each region or segment is identified by exactly one of the Hox transcription factors.[95]

Therefore, in *Drosophila*, in each segment most of the Hox genes are silenced. In this silencing process, Polycomb group (PcG) proteins play a critical role. The PcG proteins organize in 2 mayor complexes, called Polycomb repressive complex 1 (PRC1)[76,96] and 2 (PRC2)[23,97,98].

PRC2 consists of Enhancer of Zeste (E(z)), a SET domain containing protein responsible for histone H3 lysine 27 methylation, Suppressor of Zeste 12 (su(z)12), Extra Sexcombs (esc) and Chromatin Assembly Factor 1, p55 subunit (Nurf55).[41,99-101] In *Drosophila* this complex binds to Polycomb response elements (PREs). These are short genomic sequences characterized by their ability to confer silencing to adjacent reporter transgenes through Polycomb repression.[102] In consequence, PRC2 through the methyltransferase activity of E(z) will trimethylate histone H3 lysine 27 around the PRE including adjacent genes.[98]

PRC1 consists of Polycomb (Pc), Polyhomeotic (Ph), Posterior Sex Combs (Psc) and Sex Combs Extra (Sce).[103] It too binds to PREs to silence adjacent genes.[77,104] PRC1 ubiquitinates H2A at lysine 118 through Sce's ubiquitin ligase activity.[105] However, whether this step is necessary for silencing is controversial.[106,107] Additionally PRC1 also binds H3K27me3. Thereby PRC1 is able to compact the chromatin it locates to, consequently silencing genes located in the compacted chromatin.[77,104] In summary, 3 essential steps mediate Polycomb based gene silencing: The

binding of PRC2, the writing of the H3K27me3 mark[89] and the binding and of chromatin compaction by PRC1[77,104]

Drosophila also has a dedicated enzyme for the removal of H3K27me3 called Utx Histone Demethylase (Utx).[108,109] Like many other Hox activating enzymes it is part of the Trithorax group (TrxG) of proteins. [110,111] Interestingly, the enzymatic activity of Utx is important in early embryonic development, but zygotic protein is sufficient for animals to complete development and make it to adulthood.[112] However, the Utx mutant adults then rapidly die, indicating a role for Utx, and perhaps Polycomb, outside development (which would be completed at that stage). [112]

4.1.6 Polycomb and epigenetic memory

Most interesting about this silencing conferred by PcG proteins is that it is self-propagating, self-maintaining and even heritable across cell divisions.[113] Both PRC1 and PRC2 can bind to H3K27me3 which promotes a spreading of the complex across silenced regions.[114-116] This allows silencing to be maintained for a few cell divisions even if the PRE, the primary recruiter of PcG proteins, is lost thereby promoting the stability of the silencing.[25]

Importantly, PREs and by extension Polycomb group proteins appear to not discriminate between transcription factors with which they cooperate when silencing genes. This has been well described in the establishment of transcriptional patterns of Hox genes in the early embryo. Each Hox gene is selectively expressed along the anterior-posterior axis by a combination of enhancer and PREs. This patterning can also be transferred to reporter constructs using the same set of enhancers and PREs. Interestingly, while each set of enhancers mediate gene expression which closely matches just that of the Hox gene they derived from, PREs can be freely swapped between constructs: As long as it is sufficiently strong the PRE will maintain the transcriptional pattern of the enhancer and not superimpose the pattern of the gene from which the PRE derived.[102,117-120]

Further, Polycomb group proteins confer a long-term memory to Hox genes on which they act. While the enhancers which initially coordinate the segmental expression of Hox genes are only active for a relatively short time in the early embryo, the resulting transcriptional patterns are maintained throughout the fly's lifespan. [121,122] Also, this silencing memory is not tied to the activity of the enhancer that established it, it can act on any nearby enhancer. In one experiment it was shown that in a reporter construct Polycomb was able to remember the positional information initiated by a neighboring embryonic enhancer and transfer this information to an adjacent late larval enhancer when this enhancer became active many cell divisions later. [102]

Likewise, PREs, in some instances, can maintain active transcriptional states, a feature which is perhaps conferred by the activity of Trithorax group proteins. [121-123] This has led to the conclusion that, at least for these well studied Hox and developmental PREs, the transcriptional state that they confer is bistable, meaning that only fully active or repressive gene states are stable, while intermediates are unstable and therefore self-correct to be active or repressed.[113] Mathematical modelling has demonstrated, that at least on a histone modification level the enzymatic machinery of Polycomb and Trithorax group proteins could explain much of the bistability of the PREs.[66]

These properties have however been tested with only a small set of perhaps more extreme PREs that confer strong silencing to Hox genes and other developmental genes. Even among currently known PREs the strength with which Polycomb is able to silence adjacent genes varies broadly [124] and likely only the stronger PREs have so far been identified. This calls into question, in how far the described properties of Polycomb translate to targets with less well studied or no known genetic control elements.[113]

4.1.7 Towards a broader role of silencing by Polycomb group proteins

Concluding, there is a lot of genome wide information about the distribution of histone PTMs and chromatin modifications that correlate well with the observed gene state. This includes the broad spread of histone H3 lysine 27 trimethylation, which is set by the Polycomb repressive complex 2 and is found throughout the genome on silenced genes.[55,69]

Also, Polycomb group proteins confer stable long-term memory and repression to developmental genes wherever it is necessary and these genes are maintained active or inactive according to the patterning set during early development.[102,117-119,122]

This raises the question, how do H3K27me3 and Polycomb act on those targets which are not developmental in nature, but that are still targeted by H3K27me3? As Steffen and Ringrose [113] put it: "Hence, we propose that epigenetic memory may be an extreme case and that a more global function of PREs may be to stabilize or even amplify existing gene expression states over shorter time frames. Whether a given PRE mediates long- or short-term and large or small effects on gene expression states will depend on its genomic context, the DNA sequence of the PRE, the expression status of the gene and the developmental context."

Consequently, in this thesis I wanted to address whether the same principles that were discovered in the developmental context also apply in a system, where gene regulation needs to

be less terminal, but highly responsive to an external stimulus. For this I turned to the *Drosophila* immune system.

4.2 Drosophila immunity

4.2.1 Cellular immunity in Drosophila Melanogaster

Drosophila – like most animals – face the constant threat of falling prey to one of a number of pathogens and parasites. Therefore, like most metazoans it employs a large arsenal of immune response effectors, including phagocytosis (the engulfment of pathogens by dedicated immune cells) and humeral effectors (small molecules and peptides that directly harm intruders).

To ensure that wounds and intruding threats can be sensed and dealt with the *Drosophila* hemolymph – the primitive arthropod equivalent to blood – contains highly specialized immune cells, the hemocytes. These are migratory cells that employ innate immune strategies, participate in tissue maintenance and wound response and are therefore functionally similar to the mammalian myeloid cells.[125] There are 3 principle types of hemocytes that have been described: plasmatocytes, crystal cells and lamellocytes.

Plasmatocytes make up the majority of all hemocytes comprising around 95% of the total hemocyte cell pool in larvae under non-stress conditions.[126-128] Because plasmatocytes fulfill many of the same functions as mammalian monocytes or macrophages, they have also been called drosophila macrophages. [129,130] Just like macrophages, plasmatocytes are migratory, they can phagocytose pathogens and can transcriptionally respond to stimuli to produce antimicrobials and cytokines.[131-133] I will return to effectors of plasmatocytes and their regulation later.

Crystal cells are a relatively rare cell type, which under normal condition make up 2-5% of all hemocytes in larvae.[134,135] Animals lacking crystal cells are less efficient in wound healing and are more susceptible to dying after wounding[136-138], likely because of the importance of crystal cells in melanization after injury.[139,140] Crystal cells express two forms of Prophenoloxidase, PPO1 and PPO2.[141] While PPO1 is readily released into the hemolymph, PPO2 is stored inside crystalline inclusions which give the crystal cell its name and is released in a later stage of immune activation.[139] When prophenoloxidase is released and converted into active phenoloxidase these enzymes will trigger melanization, the polymerization of melanin.[142,143] Animals deficient in crystal cells or PPO1 and PPO2 are also more susceptible to infection, potentially through the lack of reactive oxygen species (ROS) normally produced as part of the melanization cascade.[139,140,144]

Lamellocytes are a highly specialized type of immune cells whose primary effector is encapsulation, the enclosing of dead tissue and parasites too large to be phagocytosed. They are larger than other hemocytes, flat and disc shaped.[145] In normal healthy animals lamellocytes are absent. They are produced in wild-type animals after wasp parasitization, where when the

parasitic wasp lays its egg into a drosophila larva, this egg is recognized by plasmatocytes which attach to it.[146] Consequently lamellocytes can be produced two different ways: On one hand, signaling through JAK/STAT, JNK and other signaling pathways[147-150] can trigger proliferation and differentiation of hemocytes in the lymph gland, a hematopoietic organ in *Drosophila* larvae, to specifically produce and release lamellocytes.[134,150] On the other hand the plasmatocytes bound to the egg can transdifferentiate to lamellocytes.[151] Lamellocytes deriving from either source can then encapsulate the egg separating it from the host, preventing growth or spread and eventually killing it though melanization orchestrated by lamellocytes and crystal cells.[146]

4.2.2 HEMATOPOIESIS IN *DROSOPHILA*

Before turning back to plasmatocyte functions, I want to look at the developmental origins of hemocytes. Three major proliferative hubs have been described for the production of hemocytes: the embryo, the lymph gland and larval hematopoietic hubs.

The earliest hemocytes are produced as early as stage 7 of embryonic development.[152-154] Estimates have put the number of cells produced at this stage at 600-700, and while a small number of the generated hemocytes are crystal cells, plasmatocytes comprise around 95% of embryonic hemocytes.[127,152]

Around the same time of embryonic development, a group of cells in the dorsal mesoderm separates. These will eventually become the dorsal vessel and the lymph gland.[155] The lymph gland is perhaps the best described hematopoietic organ in drosophila, potentially because of its large size and the number of cells produced. It consists of three major zones: The posterior signaling center (PSC), the medullary zone (MZ) and the cortical zone (CZ). The CZ is the outermost part of the lymph gland where almost mature hemocytes reside, most of them being plasmatocytes and few crystal cells and lamellocytes.[156,157] The MZ contains hemocyte progenitors which are actively dividing, maintained by proliferative signaling from the PSC,[156,158] This complex signaling network maintains the lymph gland and it keeps increasing in cell count until the late 3rd instar stages of the larval development[159-161] eventually reaching 3000-5000 hemocytes.[134,156,162]

In parallel, the embryonic hemocytes establish additional niches, called hematopoietic pockets, between the epidermis and the muscle layer of the larvae. These cells proliferate throughout the larval development while they are kept adherent and proliferative by the cells of the local microenvironment and local sensory neurons.[162-164]

In non-stress situations, both the lymph gland and the hematopoietic pockets proliferate while releasing relatively few cells until the late 3rd instar larval stages. Around the wandering stage of

the 3rd instar larvae this changes and both hematopoietic niches break down releasing large numbers of hemocytes.[162,165] While the relative contributions of both hematopoietic niches remains unclear, the hemocyte pool increases in this way from several hundred after embryogenesis to more than 5000 in early pupae stages[134,163,165], reducing to 1000-2000 hemocytes in adults.[134] Adult hematopoiesis is still debated but is most likely not massively significant.[128,166]

It is important that, even though plasmatocytes can derive from all these different hematopoietic sources, no clear distinction has yet been made that would indicate that plasmatocytes from these different sources differ in their effector functions. While many markers exist that enable the genetic marking hemocytes and even though these markers are not expressed with complete penetrance across all plasmatocytes, no subset of plasmatocytes has been described, that differs not in their origin, but in their ability to phagocytose, to respond to wound healing or in their general transcriptional programming.[135,152,162,167,168]

4.2.3 Plasmatocyte effector functions

Now I want to turn back to plasmatocytes and look at how they deal with pathogens, with emphasis on phagocytosis, wound healing, detection of pathogens and the resulting transcriptional changes.

The fact that plasmatocytes are the professional phagocyte, both responsible for the clearance of dead cells and pathogens has contributed to them being called *Drosophila* macrophages.[130] Just like in the vertebrate system phagocytosis is controlled by a set of specialized receptors and enhanced by opsonization. In development, because *Drosophila* is a holometabolous insect there is not only tissue modeling in the embryo but also tissue destruction and remodeling in pupae. And of course phagocytosis by plasmatocytes plays an important role in both embryonic development[152,169,170] and pupal development.[171-173] These developmental phagocytic processes are initiated by the dedicated receptors Croquemort [174], Draper[175] and Nimrod C4.[176]

In pathogen clearance there is a different set of receptors for phagocytosis of bacteria and fungi. Both Eater[133] and Nimrod C1 [168] are required for phagocytosis in *Drosophila* larvae and adults, and loss of these receptors leads to an increased susceptibility to infections.[133,168] Interestingly Peptidoglycan Recognition Protein LC (PGRP-LC), which also has a role in NF-κB signaling, is also a phagocytic receptor.[177] Plasmatocytes also express a set of C3/C4/C5-complement-like proteins called Thioester-Containing Proteins (TEPs) which opsonize bacteria and fungi. This opsonization further enhances the ability of plasmatocytes to phagocytose the opsonized pathogens.[141,178,179]

Plasmatocytes and crystal cells are also critical in injury response and wound healing. In the embryo any damage to the epidermis leads to an immediate response of nearby cells. First, a calcium wave is released and propagates from the site of infection.[180] This causes the NADPH oxidase, Doux, to produce H_2O_2 which is sensed by hemocytes and attracts them to the site of injury.[180,181] Similar processes guide the hemocyte dependent wound healing in larvae. Here, plasmatocytes also phagocytose damaged or dead cells and help in wound clotting.[182-184] Hemocytes further contribute to wound clotting by the production of a number of secreted protein factors. One of these is Prophenoloxidase, which, as already discussed, helps melanization based clotting.[142,143] Hemocytes also produce Hemolectin, which is a von Willebrand factor-like clotting protein.[182,185] Further, Tiggrin, an extracellular matrix protein and integrin ligand is produced by hemocytes and involved in wound healing.[185]

Besides their function in immunity hemocytes do also play a critical role in development. In addition to their role in the clearance of apoptotic cells through phagocytosis in the embryo[152] and pupae[173] that I already discussed, they also are important in extracellular matrix deposition and maintenance in embryo[186], larvae[187] and pupae[171]. There they facilitate the formation of the extracellular matrix and production of Collagen IV[186], Laminin[188], Triggrin[189] and more, and disruption of such processes have been shown to lead to developmental defects in neurons[190] and the renal tubules.[191]

Any activation either through phagocytosis, signaling from wounding or direct sensing of pathogens also triggers a transcriptional response in plasmatocytes. Among these genes, which are dynamically induced after stress or bacterial challenge are those coding for the Thioester-Containing Proteins.[141,178,179] Additionally, hemocytes, like all immune-competent tissues, produce anti-microbial peptides (AMPs), whose expression is controlled by NF-κB signaling and are specific to the pathogen.[141,192] Plasmatocytes have also been reported to secrete cytokines, even though this field is far less well studied then the vertebrate system. Unpaired signaling by the proteins Unpaired1, Unpaired2 and Unpaired3, which activate JAK/STAT pathways through their receptor Domeless and which are closely related to vertebrate leptins, have been demonstrated to play a role in drosophila immunity.[193] Unpaired3 for example is induced after infection, it signals to the fat body, *Drosophila's* liver-like organ, where it induces a stress response.[194] Unpaired3 has also been demonstrated to be important in the response to wasp parasitation.[164] Eiger is a tumor necrosis factor (TNF) like protein[195] which is upregulated after infection.[192] It plays a role in the clearance of extracellular pathogens and mutants show reduced survival, are less phagocytic and produce less AMPs in response to such challenges.[196,197] Lastly, I want to mention the cytokine like factor Spätzle, which is also expressed in hemocytes and

hemolymph and is secreted in a pro-form.[131,198,199] Mutants in *spätzle* are more susceptible to fungal infections.[131] After immune challenge by a pathogen Spätzle is converted to its active form through cleavage by the Spätzle Processing Enzyme, which is also produced by hemocytes.[199] It then binds to its receptor Toll which in turn activates NF-κB signaling[200].

4.2.4 NF-κB SIGNALING PATHWAYS IN *DROSOPHILA*

Toll signaling is one of 2 NF-κB signaling pathways that have been described in *Drosophila*.[201] In contrast to the mammalian system, where the Toll Like Receptors (TLRs) are pattern recognition receptors,[202,203] Toll in *Drosophila* does not directly bind pathogen associated molecular patterns (PAMPs).[204,205] *Drosophila* Toll signaling was discovered in the embryonic development, where its role is not that of protecting against pathogens, but instead is important in the dorsal ventral patterning.[206] During embryogenesis the extracellular protein Spätzle is cleaved by the protease Easter from its pro form into its active form and then bind to its receptor Toll. This signaling defines the dorso-ventral polarity of the embryo.[198,207,208] In immunity, extracellular cleavage of Spätzle also proceeds *Toll* signaling, but the actors involved in cleaving it differ. Here the cascade is activated by extracellular receptors recognizing either peptidoglycan from gram-positive bacteria, through the Peptidoglycan Recognition Proteins (PGRPs)[209,210] or fungal β-glucans, through the Glucan Binding Proteins (GNBPs).[211-213] This binding leads to the activation of the Spätzle Processing Enzyme, which then in turn cleaves pro- Spätzle into Spätzle.[204] Just like in development active Spätzle will then bind to Toll, which leads to its dimerization.[205,214] The subsequent intracellular signaling cascade closely resembles the pathways seen in the mammalian system.[201,202,215] The dimeric Toll receptor is bound by MyD88 with its intracellular Toll/interleukin-1 receptor (TIR) domain, which recruits a larger protein complex. This culminates in the phosphorylation of the protein Cactus, its subsequent ubiquitination and degradation.[216-218] This releases the NF-κB like transcription factors Dorsal and Dorsal-related immunity factor (DIF), which allows their translocation to the nucleus and activation of NF-κB target genes.[219,220] The other *Drosophila* NF-κB signaling pathway is called Immune Deficiency (Imd) pathway. Here the cell surface receptor Peptidoglycan Recognition Protein LC (PGRP-LC), which is also a receptor of phagocytosis, directly recognizes the peptidoglycan of gram-negative bacteria.[177,221-223] This causes the intracellular binding of the namesake of the pathway, IMD, to PGRP-LC.[224,225] IMD then recruits FADD and the caspase-8 homologue DREDD to the receptor complex.[226,227] This membrane bound complex activates the TAK1/TAB2 complex[228-230], which in turn signals through Ικκ signalosome. Both the Imd-FADD-DREDD complex and the Ικκ signalosome then

contribute to the cleavage of Relish, which releases its active N-terminal part to translocate to the nucleus, act as NF-κB like transcription factor and activate immune genes.[227,231]

Initially, a high pathway specificity for the transcription of certain AMPs in response to different immune challenges was described. It was observed for example that the AMP Drosomycin was expressed predominantly in response to gram-positive bacteria and fungi, while for example Diptericin was expressed in response to gram-negative bacteria.[232] It was therefore thought that the Toll and Imd pathways could control the expression of one or the other target through its unique NF-κB protein. Indeed, bioinformatic analysis could identify some variants of the NF-κB motive which would bind Relish or Dorsal/DIF with increased affinity, but binding of the alternative site was still high.[233] It therefore remains unclear, whether this distinction in signaling is entirely dependent on the NF-κB pathways, or whether further signaling pathways are necessary to distinguish between different types of pathogens.

4.2.5 Gene regulation by Histone Modifications in immunity

The distribution and relevance of histone posttranslational modifications has not been extensively studied in *Drosophila* immunity. To my knowledge, no genome-wide data exists on the presence of any histone marks of any drosophila immune tissue. One of the few reports dealing with epigenetic regulation in immunity described the role of the transcriptional regulator Akirin. It was demonstrated to interact with Relish after Imd signaling which lead to the recruitment of the Switch/Sucrose Non-Fermentable (SWI/SNF) complex, a nucleosome remodeling complex.[234] Knock-down of this interaction cascade lead to reduced $H3K4ac$ after immune challenge, reduced expression of immune genes after immune challenge and an overall drop in survival of infected animals.[235]

More reports are available on histone modifications in mammalian macrophages and on the role of $H3K27me3$ in mammalian innate immunity. The availability of large numbers of monocyte or bone marrow derived macrophages has made them a useful tool for the study of cellular processes in primary cells. Therefore, many genome wide profiles are available from such cells.[236]

A number of reports show the importance of epigenetic regulation in the differentiation of innate immune cells, demonstrating the requirement of PcG and TrxG proteins in the differentiation of monocytes to macrophages[237] as well as in the plasticity of $CD4^+$ T-cells.[238] However, some data is also available on the role of $H3K27me3$ in gene regulation in innate immune cells. Two reports demonstrate a potential role of two histone H3 lysine 27 demethylases KDM6A and KDM6B, which are orthologs of the *Drosophila* Utx Histone Demethylase. One report demonstrates, that knock-down of KDM6A increases the production of the cytokines IL6 and IFN-β in response to

viruses and TLR activation.[239] Another report showed, that KDM6B is strongly induced after LPS stimulation of macrophages, but they did not investigate how this affected gene expression.[240] Histone modifications have also been implicated in immune memory, the ability of immune cells to respond differently to a stimulus when encountering it for a second time, either less or more severe than the first time around. It was demonstrated that H3 lysine 27 is differentially acetylated after β-glucan or LPS priming and might therefore play a role in immune memory.[241] Another study showed, that H3K27me3 was induced at a small set of promotors after stimulation with the cytokine IFN-γ, thereby acting anti-inflammatory.[242]

I think that studying histone modifications in immunity is a good complementation of the long-established study of those modifications in development. Immunity has some interesting features of gene regulation, that set it apart from development: Genes must be dynamically regulated, they should be transcriptionally silent in non-immune challenged situations, but must be induced rapidly after infection. The central aim of my thesis was to characterize how the regulation of those genes, which are just like developmental genes often silenced, but whose silencing can be overcome quickly, is implemented on the chromatin level. I did so by studying histone modifications in general and H3K27me3 in particular in *Drosophila* plasmatocytes and I hope that the information gained from studying such a system might also be translatable to similar, highly responsive signaling situations, such as neuronal gene regulation or metabolism.

5 RESULTS

5.1 Transcriptional response of Drosophila hemocytes to infection

5.1.1 THE SEPTIC INJURY MODEL IN *DROSOPHILA* LARVAE

The septic injury model is a well-established model to study the immune response in *Drosophila*.[243] In this model bacteria are introduced into the animal by pricking it with a sharpened tungsten needle covered in bacteria, which can be individual strains or mixtures of different bacteria and is not limited to *Drosophila* pathogens.

To demonstrate that both the Imd and Toll pathways are activated by the manipulation I aimed to reproduce the available findings on the activation of those pathways by septic injury in larvae.[232] Therefore, I tested wild-type Oregon R (OrR) 3rd instar wandering larvae in their response to the microbial challenge to the gram-negative bacterium *Escherichia coli* (*E. coli* or E.c.), the gram-positive *Staphylococcus aureus* (*S. aureus* or S.a.) and to sterile injury. 6 hours after the injury I isolated total RNA from whole larvae and subjected it to RT-qPCR determining expression levels of the AMP genes *Diptericin* (*Dpt*), which is a target of the Imd pathway, and *Drosomycin* (*Drs*), which is regulated by the Toll pathway.[232] My results demonstrate that in my hands, as previously reported, the expression of both AMPs is increased in all 3 treatment conditions (*E. coli*, *S. aureus*, Sterile injury, Figure 1A). Further, the response of *Dpt* is highest in the *E. coli* challenge, consistent with a predominant activation of the Imd pathway, and the response of *Drs* is highest in the *S. aureus* injury, indicating the preferential activation of the Toll pathway.

To further characterize the transcriptional response of larvae to septic injury I challenged mutants of NF-κB signaling pathways and measured their response. First, I tested a mutant deficient in the Imd pathway with a loss of function mutation in the NF-κB transcription factor *Relish* (*Rel*).[244] I performed RT-qPCR against the AMPs *Dpt* and *Drs*, which are reporters of the activation of the Imd and Toll pathway, respectively (Figure 1B). In the Rel^{E20} mutant, the expression of *Dpt* is strongly reduced in response to sterile injury, *S. aureus* challenge, and *E. coli* challenge. The response of *Drs* was also reduced following all 3 challenges, though to a lower extent. This corresponds well with previously published data.[232,244]

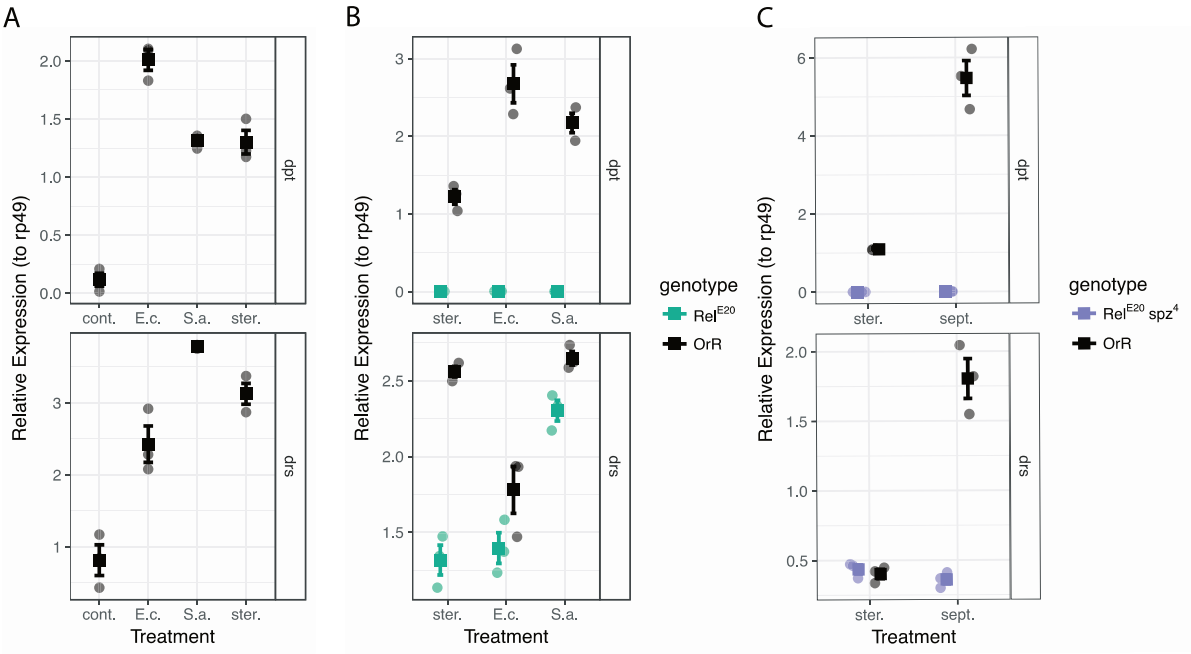

Figure 1: Septic injury induces an NF-κB immune response in *Drosophila* larvae. A: Wild type OrR larvae were pricked with a sharpened tungsten needle, either sterilized with ethanol (sterile injury) or dipped into a pellet of *E. coli*, *S. aureus*, or kept untreated (cont.). Total RNA was isolated from these larvae and was subjected to RT-qPCR. Ct values of *Drs* and *Dpt* were normalized to rp49. One-way ANOVA p-value: *Drs* = $1.62*10^{-5}$, *Dpt* = $7.85*10^{-7}$. B and C: Larvae from mutant background Rel^{E20} (B) or Rel^{E20},spz^{4} (C) were pricked with a sharpened tungsten needle, either sterilized with ethanol (ster.) or dipped into a pellet of *E. coli* (E.c.) or *S. aureus* (S.a.), or a mixture of *E. coli* and *S. aureus* (sept.). Total RNA was isolated from the larvae, RT-qPCR performed and the data was normalized as in A.

Next, I tested Rel^{E20} spz^{4} double mutant larvae that lack signaling in both NF-κB pathways.[226] In these experiments I challenged 3rd instar wandering larvae with sterile injury or a septic injury (a combination of *E. coli* and *S. aureus*) and performed RT-qPCR on *Dpt* and *Drs*. The results indicate that both, expression of *Dpt* and *Drs*, were not induced by the challenge. (Figure 1C)

Interestingly, while the requirement of the Imd pathway to induce *Dpt* and the requirement of *spz* to induce *Drs* was be recapitulated very well in this work, the specificity of the pathways to *E. coli* and *S. aureus* only reproduces partially. While the first experiment (Figure 1A) shows, as expected the highest response of *Drs* to *S. aureus* and the highest response of *Dpt* to *E. coli*, the second experiment has mixed results with regard to PAMP specificity (Figure 1B and C). Therefore, it is possible that other factors contribute to the transcription of specific AMPs, that has not been addressed currently.

My data and the corresponding literature show, that injury alone will also cause transcriptional changes in the animal and its hemocytes[141,243]. In this thesis, however, I am interested in the

general mechanism of transcriptional regulation and therefore the relative contributions of damage-associated molecular patterns (DAMP), pathogen-associated molecular patterns (PAMP) and wounding response to the transcriptional response are not under investigation. Instead, I am aiming to achieve a maximally strong transcriptional immune response. Hence, I decided to apply a septic injury model that combines a *E. coli* and *S. aureus* immune challenge.

5.1.2 PLASMATOCYTE ISOLATION AND PURITY

To study the plasmatocyte specific response I, therefore, adapted a method to isolate these cells from *Drosophila*. This resulted in a slightly modified version of the published protocols[245] that I used throughout my thesis. In brief, 3rd instar wandering larvae were used for plasmatocyte isolation, because they have the highest hemocyte count and can be bled while avoiding the disruption of the gut or other organs. These larvae were bled out into a tissue culture dish containing Schneider's medium as described in the methods. I also added the ROS scavenger N-Acetyl-L-Cysteine (NAC) to block the melanization[246] that disruption of the cuticula will trigger. The extracted hemocytes attach strongly to the tissue culture surface and can be washed rigorously with PBS to remove debris, bacteria and potentially remaining crystal cells.

To test the purity of the resulting Plasmatocyte preparation I used a *Drosophila* line with the genetically encoded marker Hml-dsRed. In this line, plasmatocytes and crystal cells are marked by nuclear expression of dsRed.[247] I isolated hemocytes from larvae of this line and the wild type OrR strain and analyzed them by flow cytometry to determine if residual non-fluorescent cells are present in my preparation (Figure 2A-C).

Plasmatocytes were first gated by forward scatter (FSC) and side scatter (SSC) to select intact cells. Considerable subcellular debris was present in the preparation (Figure 2A), likely due to the forceful method by which the plasmatocytes are extracted. However, A clear population of intact cells was identified and selected. The gated single cells (Figure 2B) showed a strong separation between OrR and Hml-dsRed cells, with minimal overlap between plasmatocytes isolated from the non-fluorescent wild type strain OrR and the Hml-dsRed line (Figure 2C). This indicates that the gated cells were exclusively hemocytes.

However, both plasmatocytes and crystal cells express this Hml-dsRed reporter[126], so these two cell types are not distinguishable by this method. Therefore, I checked cells isolated from the Hml-dsRed line by fluorescence microscopy, where crystal cells can be identified by their distinct internal morphology and size.[248] Consistent with flow cytometry, all cells showed dsRed expression with distinct nuclear localization and morphologically these cells were homogeneous

with no crystal cells detectable (Figure 2D). Together this demonstrates, that my adaption of the hemocyte isolation protocol works and produces a pure and homogeneous population of plasmatocytes.

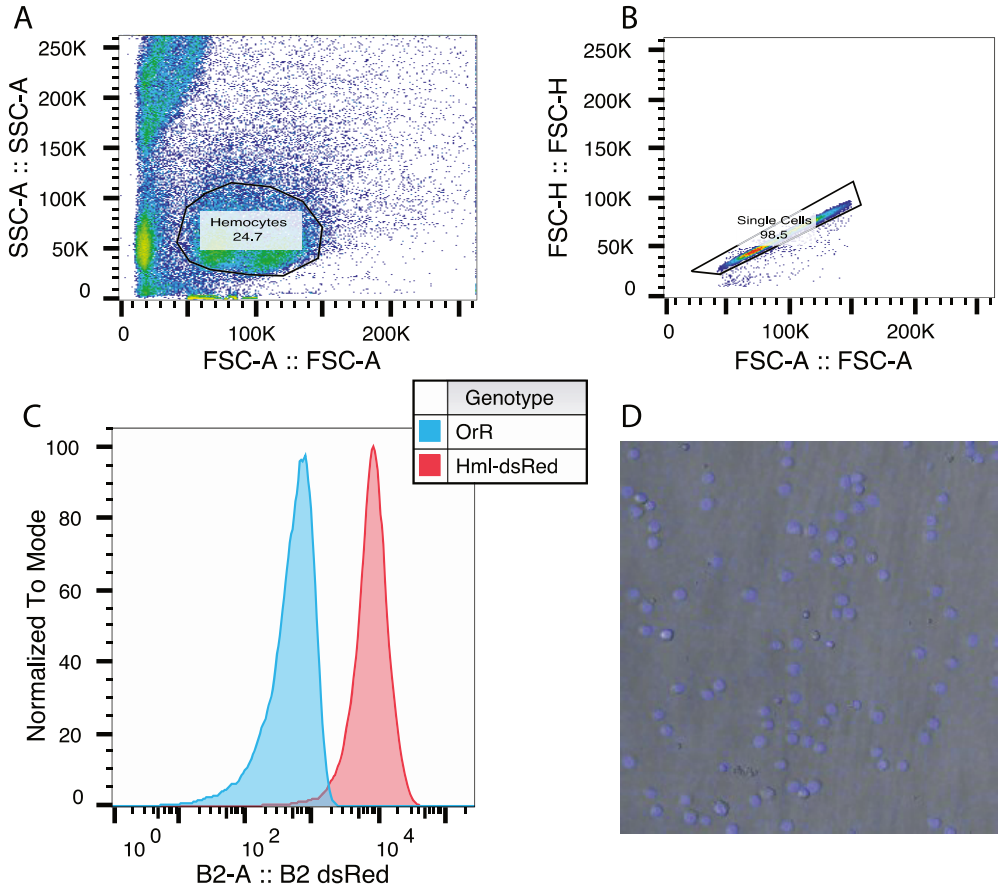

Figure 2: Plasmatocytes can be isolated readily from *Drosophila* larvae. A – C: Plasmatocytes were isolated from OrR and Hml-dsRed lines and FACS sorted. A: FSC vs SSC of all isolated cells from the Hml-dsRed line. The gate marked "Hemocytes" marks the selection of events that are considered intact cells. B: FSC height vs area of all cells selected by the gate in A. Events with a high correlation in FSC height and area are considered single cells and selected. C: Isolated cells from OrR and Hml-dsRed lines were gated according to A and B. dsRed was exited with a 488nm laser and detected a 585/40nm bandpass filter. D: Hemocytes were isolated as described before but left to attach to a tissue culture dish and imaged by fluorescence microscopy. Overlay of the channels from transmission and fluorescence from excitation 531/40 nm and detection 593/40 nm.

5.1.3 RNA-SEQ TIME COURSE OF PLASMATOCYTES AFTER INFECTION

To identify immune genes in plasmatocytes I then characterized their transcriptional response at various time points after immune challenge. Accordingly, I used the septic injury model established in 5.1.1 and challenged larvae with a combination of *S. aureus* and *E. coli*. Previous

reports indicated, that the immune response in whole animals is maximal at about 6 hours post-challenge.[232] In order to determine the dynamics of the transcriptional response, I decided to also include time points at 3 and 18 hours post-challenge. For controls, I included plasmatocytes from unchallenged animals and from 6 hours post sterile injury. In order to minimalize perturbance by factors like larval density, animal age and circadian rhythm, the closely staged 1st instar larvae were seeded into food tubes at fixed densities at fixed times, and plasmatocytes were also isolated from late 3rd instar larvae at fixed times. The different time points post infections were achieved by staging the time of immune challenge 3, 6 or 18 hours before the fixed plasmatocyte isolation time. From the challenged larvae samples of plasmatocytes were collected, RNA was isolated and subjected to RNA-seq. Four independent biological replicates were collected in this fashion for every treatment condition. The resulting data was mapped and treated as described in the methods section.

First, I performed a principal component analysis (PCA) on the 1000 most variable genes in the transcriptome of challenged plasmatocytes acquired by RNA-seq (Figure 3A). In this analysis RNA samples cluster closely together according to the treatment condition from which they derive, indicating a high degree of reproducibility between biological replicates. The samples from unchallenged (control) plasmatocytes, 3h post septic injury and 6h post septic injury clearly cluster apart, and interestingly they are located on different tips of an imaginary triangle rather than being points on a straight line. This suggests that plasmatocytes do not initiate a singular transcriptional response that merely increases or decreases over time but in fact induce two distinct transcriptional programs: An early response detected after 3 hours and a late response detected by 6 hours. Almost half (49.71%) of the variance across all samples is explained by the first component of the PCA alone, and the samples from unchallenged and 3h post septic injury locate furthest apart on that scale, indicating that these samples are most dissimilar and the largest changes are to be expected between them. The other samples from challenged plasmatocytes (6h post septic injury, 6h post sterile injury, 18h post septic injury) also cluster slightly apart from the control on this dimension, but to a far lesser extent. This indicates that the early response that was observed at 3h post septic injury is much reduced at later time points. Component 2 of the PCA did not show a strong separation according to conditions. There is only a tendency for all challenged conditions to slightly separate from the control in this dimension. However, I could not find any interfering variables like date of sample preparation, order of sample preparation, RNA yield, RNA integrity or technical parameters of the RNA-seq library preparation and sequencing that explains the variance across this or any other dimension. Therefore, it might simply represent true biological variance that needs to be expected from

biological systems. Dimension 3 explains 10.2% of all variance that was observed and strongly separates the control from both the 6h post septic injury and the 6h post sterile injury. Interestingly, the 3h sample does only vary very little across this dimension, indicating that this is a distinct late response that is different from that observed at 3h post septic injury. The 18h sample is positioned in-between the 6h samples and the control, which indicates a return to the ground state after clearance of the bacteria. Dimensions 4 and 5 explain 7.45% and 2.83% of the variance respectively but do not yield any biological information and were, therefore, not shown. Another interesting observation is the overlap of the samples from 6h post sterile injury and 6h post septic injury. These plasmatocyte samples seem to be very similar, even though my initial whole larvae experiments at the same time points (Figure 1A) demonstrate that the septic injury induces an overall stronger response than sterile injury in whole larvae. It is possible, however, that plasmatocytes respond transcriptionally different than the whole animal, where the predominant source of AMP transcripts will be the fat body.[131] Plasmatocytes migrate to the wound, where they may encounter and respond to bacteria that leak in from the wound even after injury with a sterile needle. Alternatively, they may mount an anticipatory response at 6h post-challenge. In summary, at 6h post septic injury plasmatocytes seem to distinguish little between the high bacterial load introduced by septic injury and an injury with a sterile needle.

The findings from the PCA can be corroborated by identifying differentially expressed genes. For this, I filtered for genes that are represented with at least 1 read mapped to them per 1 million total reads in any 4 samples, regardless of which treatment group they belong to. This left me with data for 7586 out of the 17737 genes annotated in BDGP *Drosophila* genome assembly 6.91. Filtering genes in such a fashion is useful because changes in genes with lower read count are not likely to reach significance, but their presence during testing increases p-value thresholds during false discovery rate correction. Therefore, I will apply this filtering in all RNA-seq experiments in this thesis before testing for differential expression. On this gene set, I applied pair-wise comparisons of conditions using edgeRs glm approach and aimed to identify genes with a $\log_2$ fold change in expression greater than 1 (LFC > 1) and a false discovery corrected p-value smaller than 0.05. First, I compared the 6h post sterile injury and the 6h post septic injury samples. Consistent with the PCA analysis there are very few differences between the two samples, and no genes reach the significance threshold outlined before.

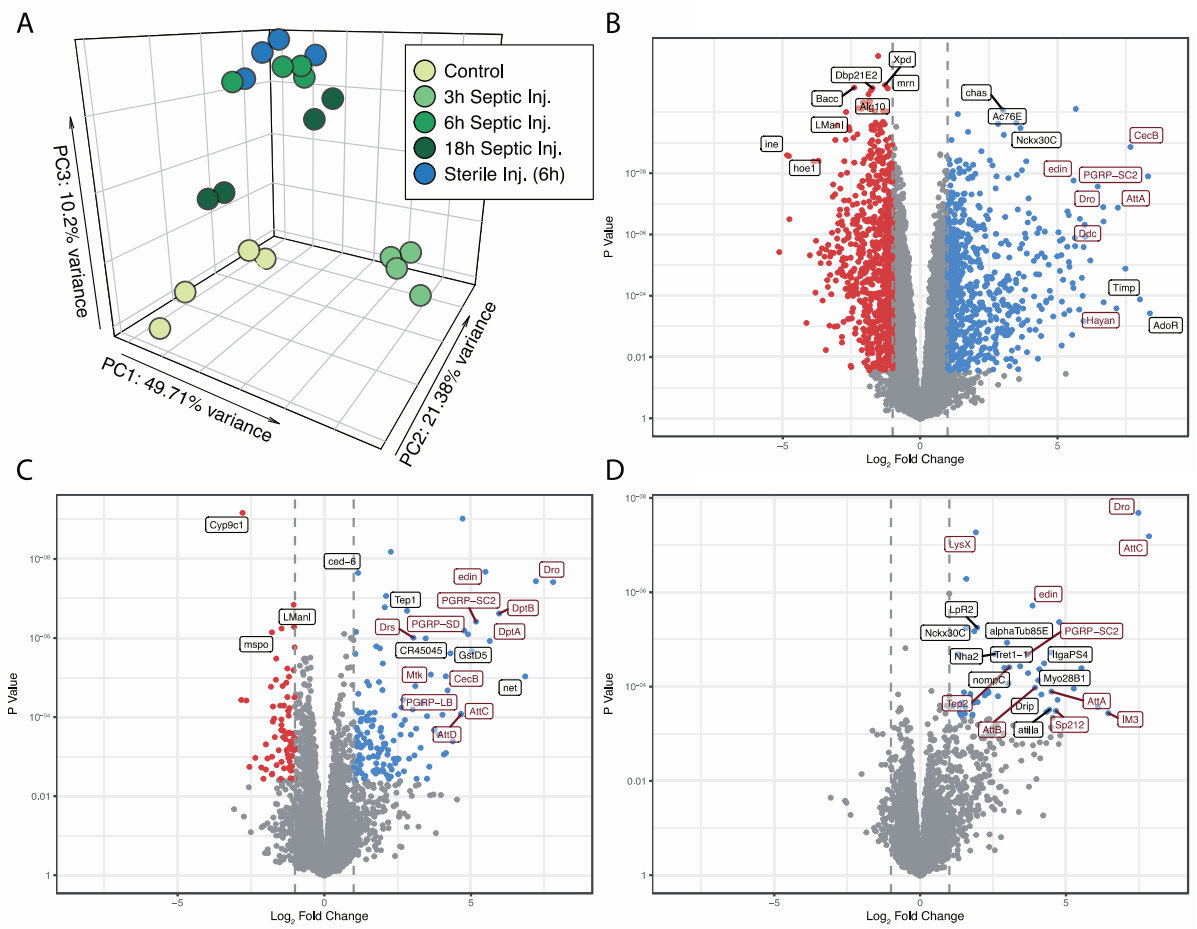

Figure 3: Septic injury induces a multi-component transcriptional response in plasmatocytes. A: PCA of RNA-seq from immune challenged plasmatocytes, septic injury indicates a combination of *S. aureus* and *E. coli*. Gene level read counts were stabilized by applying a regularized logarithmic transformation (DESeq2), then PCA was performed on the 1000 most variant genes. B-D: Volcano plots of all detected genes for the comparisons 3h post septic injury vs. control (B), 6h post septic injury vs. control (C) and 18h post septic injury vs. control (D). P-values are plotted log transform against the Log_2 fold change (LFC). P-values were calculated by a glm negative binomial approach (edgeR) and significant genes were selected with a cut-off false discovery rate corrected p-value = 0.05 and LFC = 1 (marked by vertical dashed lines). The most outlying genes are marked by gene name, and genes with immune GO terms among the labeled genes are labeled in red.

Subsequently, I determined how plasmatocytes transcriptionally responded to septic injury at various time points after the challenge. I, therefore, compared the septic injury samples from 3h, 6h, and 18h post challenge to the unchallenged control samples and determined log2 fold changes, p-values, and differential expression. At 3h after septic injury (Figure 3B), a large number of genes were differentially expressed (705 genes are up-regulated, 6098 are not significantly differentially expressed, 783 are down-regulated). While a larger number of genes were down-regulated than up-regulated, the most strongly regulated genes, with the largest fold

change, were up-regulated. Among them are many genes that are associated with at least one immune gene ontology (GO) term. Furthermore, when comparing the volcano plots from 3h post septic injury vs. control (Figure 3B) to the ones from 6h and 18h post septic injury vs. control (Figure 3C-D) it is apparent that the number of differentially expressed genes was far greater at 3h post-challenge then at any other time point. This observation is well consistent with the PCA. At the same time, the extent of up-regulation in the most strongly up-regulated genes is similar in all comparisons, hence the immune response at 6h or 18h post septic injury is not simply a weaker induction of the same response. At 6h post septic injury (Figure 3C) I detected a smaller number of regulated genes compared to the 3h time point (141 genes are up-regulated, 7373 are not significantly differentially expressed, 72 are down-regulated). Here, very few genes were down-regulated, and those that were do not show a large fold change. On the other hand, a number of genes were up-regulated with large fold changes, and among them are many genes with annotated immune GO terms (marked with red text labels in the volcano plots), specifically the AMPs of the *diptericin, attacin,* and *cecropin* groups and the receptors from the *peptidoglycan recognition protein* group. At 18h after septic injury (Figure 3D) no down-regulated genes were detected (50 genes are up-regulated, 7536 are not significantly differentially expressed, 0 are down-regulated). As for the comparisons of 3h and 6h post septic injury, a large number of immune-type genes were among the up-regulated genes. The overall smaller number of differentially expressed genes indicates that plasmatocytes at this stage were returning to a "resting" state, normally observed in unchallenged animals. However, a number of genes were still up-regulated with high significance, and their fold change was not necessarily smaller than at either 3h or 6h post-challenge. Together, this demonstrates that the transcriptional response of plasmatocytes to septic injury is a two-component response, in which early at 3h post septic injury, a large number of genes is differentially regulated, followed by a later response at 6h post septic injury, where specific immune genes are highly expressed.

To identify at the trajectory of individual genes and their reproducibility across the 4 replicates in each condition I selected a set of relevant genes by choosing the genes with the highest variance among all genes that were differentially expressed in plasmatocytes at 6h post septic injury when compared to control conditions. These genes I normalized for their mean and standard deviation of expression by applying a gene-wise z-score transformation on the regularized logarithmic read counts. These values were plotted in a heatmap (Figure 4), such that each column represents one plasmatocyte sample and each row represents a gene, sorted by hierarchical clustering. On the gene level, several clusters are discernible, indicated by colors next to the gene names: Some

genes are generally up-regulated during infection (in the upper part of the heatmap in Figure 4) and some are generally down-regulated (which are in the lower part of the same heatmap).

In the uppermost cluster is a set of genes that includes *TEP2* and the AMPs *DptB*, *Mtk* and *Drs*, which are up-regulated at 3h and 6h after the septic injury to about the same extent. Below is the cluster marked in orange which includes the *cecropins* and 2 *attacins* that is most strongly up-regulated at 3h after septic injury, but also highly expressed at 6h in septic and sterile injury.

The clusters marked in purple and magenta include genes that are more highly expressed at 6h and 18h post-challenge than in either 3h post septic injury or control. Among the genes in the yellow cluster that are generally down-regulated, there is a set of *Nimrod* phagocytosis receptors, but unfortunately, no data exist whether these specific *Nimrod* subtypes are required for developmental phagocytosis or phagocytosis of pathogens.

On the pattern of individual samples a few interesting observations can be made, both in the heatmap of differentially expressed genes (Figure 4) and the dendrogram of hierarchical clustering of those samples (Figure 5). The samples from 6h post septic injury and sterile injury are very similar, with the 3h sample being somewhat similar and the control sample being most dissimilar. The 18h post septic injury clusters in several places, which is also apparent from the heatmap, where there seems to be 2 distinct types of samples in the 18h group. Potentially some variance in how quickly the challenge was be resolved resulted in these differences. Judging by the heatmap this difference may, in fact, derive from only a subset of genes among the genes that were up-regulated at 3h and 6h, with 4 genes in the magenta cluster, among which is *AttC*, being strikingly different between the 18h samples 1 and 4 and the 18h samples 2 and 3. The genes that were down-regulated at 3h and 6h are, however, consistently returned to control levels in all the 18h samples.

Therefore, individual genes, including different AMPs, can be identified, which are most strongly regulated at different times post septic injury and thereby group the plasmatocyte samples into 3 principle transcriptional states: Unchallenged, 3h post septic injury and 6h post septic injury.

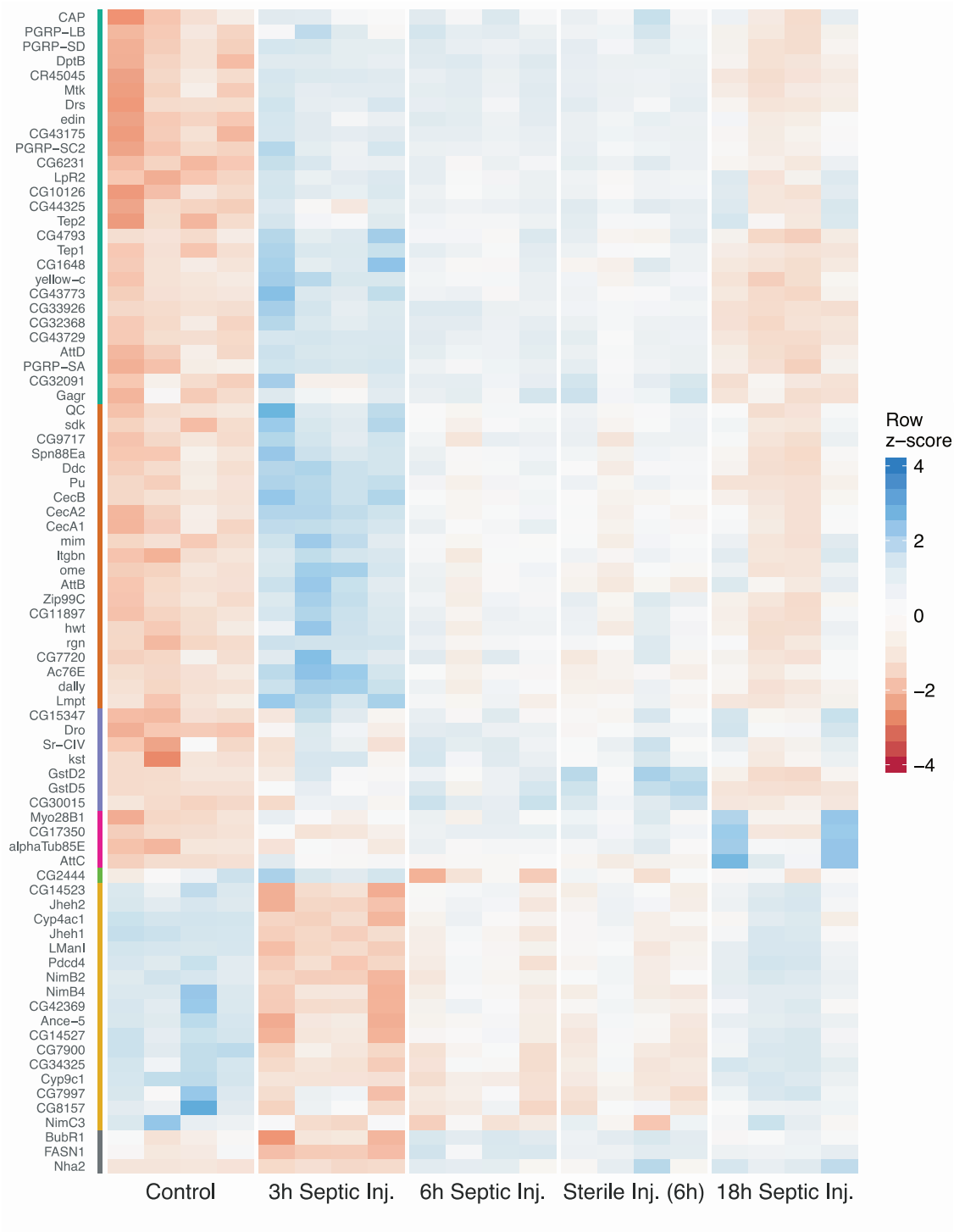

Figure 4: Plasmatocyte immune genes are regulated in a characteristic temporal fashion. Genes were selected by choosing among all genes up-regulated in 6h post septic injury the genes with the highest variance across all samples. Regularized logarithmic transformed (DESeq2) gene read counts were normalized by z-score transformation and genes were sorted by row-wise hierarchical clustering (indicated by colors on the left). For the tiles, blue indicates genes that are in that sample more highly expressed than in the mean, while red indicates gene expression lower than the mean.

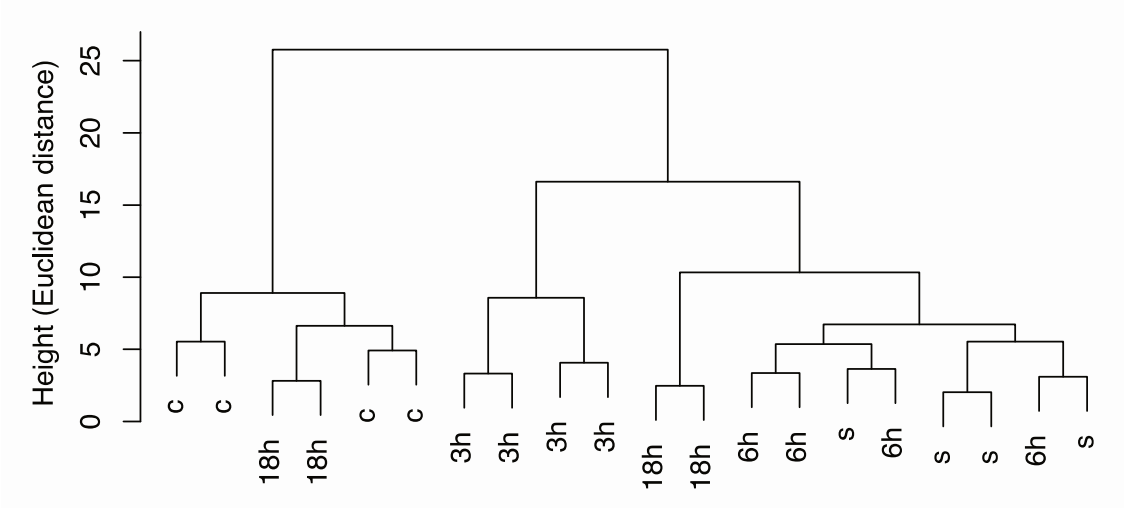

Figure 5: Plasmatocyte transcriptional immune responses cluster by their post-challenge time. The z-score transformed gene counts from Figure 4 were clustered by hierarchical clustering and plotted as a dendrogram, such that the heights represent the Euclidian distances of samples. Labels c: control, 3h: 3h post septic injury, 6h: 6h post septic injury, s: 6h post sterile injury, 18h: 18h septic injury.

To better understand the functional consequences these transcriptional changes have for plasmatocyte function, I identified the enrichment of gene ontology (GO) terms by applying the algorithm from goseq as implemented in edgeRs goana module.

GO-terms relating to immunity are enriched at 3h and 6h after septic injury but only some remain enriched by 18h after septic injury. This observation is consistent with an immune response of plasmatocytes that returns to a 'resting' state within 18h after the challenge. Generally, the effect sizes of the enrichments are higher in the 6h samples, potentially because fewer genes that are not annotated as immune effectors are differentially regulated at these time points. However, immune-related GO-terms are enriched amongst genes that are differentially expressed between 3h and 6h after septic injury. This underlines that the immune response changes from an early to a late response. Interestingly, genes that are involved in general metabolic processes and more specifically carbohydrate derivative metabolic processes are enriched among the down-regulated genes both, at 3h and 6h after septic injury. This suggests a change in the metabolic state of plasmatocytes alongside their immune response. Down-regulation of glycolysis is contradictory to the Warburg-effect observed in macrophages after infection[249], but it would be necessary to measure enzymatic activity in glycolysis to determine if and how the transcriptional changes in plasmatocytes actually affect their metabolic activity.

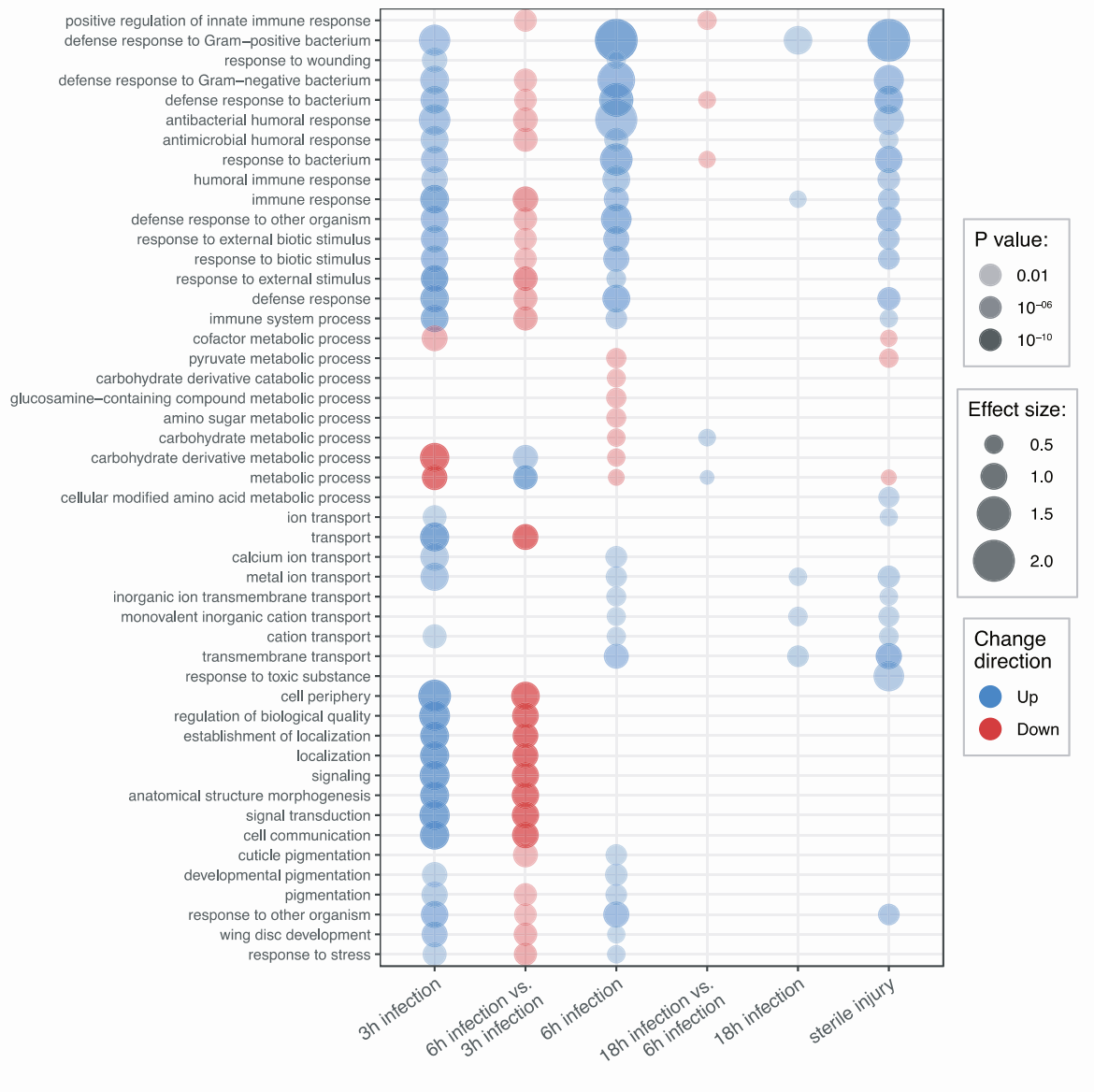

Figure 6: Specific functional programs are sequentially induced in plasmatocytes after septic injury challenge. GO-term enrichment was performed using the goana module implemented in edgeR. Rows are GO modules that reached significance, while columns are individual comparisons and, where not otherwise noted, the challenged plasmatocyte samples were compared to the RNA-seq of control plasmatocytes. The effect size is (b/n)/(B/N) where b is the number of genes that are regulated and carry the GO-term, n is the number of genes regulated in that direction, B is the total number of genes with that GO-term, and N is the total number of genes in the comparison. GO-terms are sorted by manual selection to reflect common groups of gene functions.

Further, a set of terms related to ion transport, specifically metal ions and cations, appears to be enriched among up-regulated genes at 3h, 6h, and 18h after septic injury. These could relate to calcium signaling, which is important in wound healing[180] or be related to antimicrobial activity, which can be mediated by ion transporter.[250] Further, a number of GO modules that relate to cell

to cell signaling and cell migration is enriched am genes up-regulated at 3h post-challenge. This is likely related to the signaling and the migratory response of plasmatocytes.[181] Interestingly, the genes out of this group are no longer enriched in regulated genes later than 3h post-challenge, which suggests that they are part of an early and transient response. Finally, the genes in terms relating to pigmentation, which are enriched among up-regulated genes at 3h and 6h post-challenge fulfill a dual role: Because both pigmentation and immunity use melanin, the genes contained in these terms here are likely to be related to melanization for clotting or antimicrobial ROS production.[142-144] It is important to note that also based on GO-term enrichment the septic injury condition at 6h after challenge was not readily distinguishable from injury with a sterile needle at the same time point. Still, the GO-term analysis clearly shows that plasmatocytes show a complex biological response in the septic injury model that is dominated by a pattern of immune response mechanisms.

In summary, I characterized the transcriptional response of plasmatocytes at various stages of an immune response. There are only 3 other whole transcriptome studies on *Drosophila* plasmatocytes like cells that I am aware of, all of which used microarrays. Irving, et al. [141] analyzed the transcriptional pattern in various mutants affecting hemocyte hematopoiesis, Johansson, et al. [192] tested the response of the hemocytes like mbn-2 cell line to PAMPs and Woodcock, et al. [251] investigated the effect of diet on adult hemocytes. Therefore, to my knowledge, I generated the first RNA-seq data sets on primary *Drosophila* plasmatocytes and their transcriptional regulation in immunity. Additionally, I showed that there are distinct transcriptional patterns at 3h post-challenge and 6h post-challenge, indicating that there is a two-step response, early and late. Furthermore, I showed that the response is highly reproducible and although complex, dominated by genes that have been annotated with immune functions.

5.2 ChIP-seq on unchallenged Drosophila plasmatocytes

5.2.1 ChIP-seq sample description

To establish the regulation of gene expression by histone posttranslational modifications in *Drosophila* plasmatocytes I further established a detailed profile of the relevant histone modifications targets from those cells. Therefore, I isolated plasmatocytes from unchallenged *Drosophila* larvae, prepared cross-linked chromatin from them and subjected it to chromatin immunoprecipitation (ChIP).

On average after washing, I recovered approximately 1000-2000 plasmatocytes from each larva. Given that most ChIP protocols work in excess of 1 million cells, it was not feasible to optimize ChIP protocols for each of the 10 targets in this study using plasmatocytes. Therefore, I used chromatin from the S2R+ *Drosophila* cell line, from which larger quantities of chromatin can be prepared easily, to optimize ChIP protocols. This chromatin was sheared, immunoprecipitated and enrichment was determined by qPCR for known or predicted targets (data not shown).

After optimizing my protocols to work with approximately 10^5 cells I performed ChIP experiments on chromatin from plasmatocytes for all targets. To control for chromatin accessibility, genome duplication or deletions and for potential problems in the data processing pipeline, I held a matching input sample back whenever a ChIP experiment was performed. As is standard in ChIP-seq[252], 2 replicates were prepared for each ChIP target, the immunoprecipitated DNA was tested by qPCR for enrichment in predicted regions and consequently made into libraries and sequenced (see Table 9 for libraries). The resulting sequencing reads were mapped to the Drosophila BDGP dm6 genome assembly and treated as described in the methods.

To visualize the processed data in a genome browser, I merged the replicates and calculated bigwig coverage files using 25 bp bins. An example of the resulting profiles is shown in Figure 7. The genomic region shown here is the area around the *even skipped* (*eve*) locus. The *eve* locus is controlled by a set of enhancers, that mediate a highly specific expression pattern in the embryo[253] and is silent in larval plasmatocytes, according to the expression analysis in this study. In Figure 7 I colored the tracks according to the correlation that has been ascribed to the corresponding histone modification: Modifications that mark active chromatin (active marks) in green and modifications that mark repressed chromatin (repressive marks) in red. While many regions carry some active marks, the chromatin region around the *eve* locus is devoid of all of them. In this genomic region, the repressive marks H3K9me3 and H3K27me3 as well as the 'combined' H3K27me2/me3 mark (detected by using an antibody recognizing both H3K27me2 and H3K27me3) show enrichment. Interestingly, it appears that the input in these regions is slightly reduced, but it is still well detected. For genes outside the *eve* domain, there are a number

of peaks and short regions enriched for active marks. First of all, RNA polymerase II has a strong peak as well as a trailing signal along the gene body on genes that are transcribed. This has been described as its characteristic distribution, where the initial peak indicates stalled RNA Polymerase II and the tail reflects progressing transcription.[254]

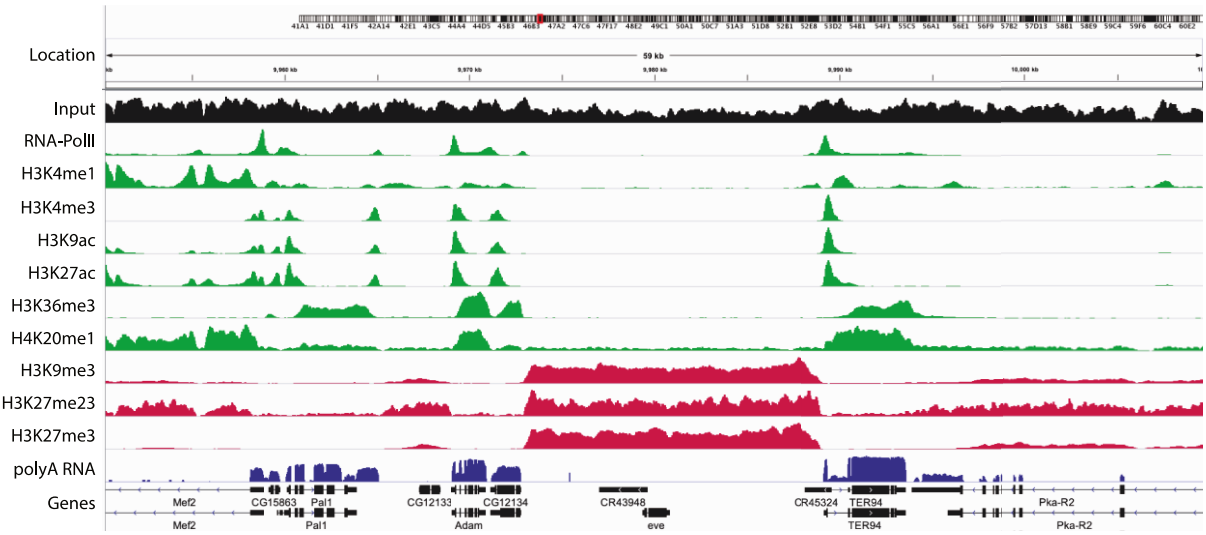

Figure 7: Genome profiles of ChIP-seq from unchallenged plasmatocytes are concordant with previously published observations. Mapped ChIP-seq reads from 2 replicates for each IP target were mapped on dm6.91 and ChIP coverage was calculated using 25 base pair bins. The resulting profiles were visualized using IGV. Here, I show the profiles around the eve locus. Green tracks label histone marks that are associated with active transcription, while red tracks label repressive marks. Additionally, the blue profile shows $\log_2$ coverage of RNA-seq from unchallenged plasmatocytes. The bottom panel shows gene spans of genes annotated in the genome assembly dm6.

H3K4me3 peaks are located directly after the transcriptional start site (TSS) of active genes, centered slightly after the stalled RNA polymerase II peak, a distribution which is well described for H3K4me3.[54] The competing H3K4me1, on the other hand, is located at positions further away from these H3K4me3 TSS peaks and also locates to introns of active genes. It is mostly restricted to active genes and the literature suggests that these regions act as enhancers.[22] The lysine acetylations H3K9ac and H3K27ac appear to be enriched on both H3K4me1 and H3K4me3 positive regions, but while H3K9ac seems to be preferentially present at H3K4me3 peaks, H3K27ac seems to be more evenly enriched among both of the tested H3 lysine 4 marks. H3K36me3 is enriched on active genes at regions downstream of the TSS while being absent from long introns like the one in *Mef2*, which corresponds well with the described enrichment for H3K36me3 at the 3' exons of active genes.[20] H4K20me1 is present along the complete gene body of active genes both on introns and exons. However, the background in this ChIP is higher than

in other ChIPs, with a signal being visible even in the silenced *eve* locus. In short, the ChIP-seq profiles show a distribution concordant with previous publications[55], which indicates that modified genomic regions were successfully enriched.

Additionally, I controlled that signals that I find for my ChIP-seq experiments reproduce. Therefore, I checked the correlation between the ChIP replicate pairs. To do this I followed the standard Deeptools pipeline[255]: The genome was divided into 10 kb bins and read counts were quantified for those bins. Then, the Pearson correlation coefficients for those counts were calculated and plotted as a hierarchically clustered heatmap (Figure 8). The label after the ChIP target (e.g. CHxxx or Kxxx) identifies the experimental group of the ChIP (see Table 9).

Pairwise Pearson correlation coefficients show that the replicates reproduce very well, and all ChIPs have a Pearson's r between r = 0.99 and r = 1. The input samples correlate to a lesser extent, but this is likely because there is little true signal in these tracks and most differences between samples will be random fluctuations in read sampling. The Input samples from the experiment K3772 are somewhat more dissimilar to the other inputs, which might stem from the lower sequencing depth in those samples.

Beyond reproducibility within replicates, this analysis reveals correlations (or anti-correlations) in the genome-wide distribution of different histone marks and therefore allows to group them into co-occurring clusters. H4K20me1 clusters most closely with input and not well with other histone marks, an observation that could be caused by the poor ChIP enrichment achieved with the monoclonal antibody used in this sample.

In contrast, the three repressive marks H3K9me3, H3K27me2/me3 and H3K27me3 cluster well together. Interestingly, H3K27me3 does not correlate very well with H3K27me2/me3, which recapitulates the results of the raw data tracks (Figure 7). Therefore, some additional signal must derive from H3K27me2, which, based on the better correlation with the input samples and the profile I described before, appears to be spread more broadly across the genome than H3K27me3.

Similar to the repressive marks active histone marks cluster together. The acetylations H3K27ac and H3K9ac correlate very strongly, but they also correlate with the other TSS associated marks H3K4me3 and RNA PolII as well as the enhancer mark H3K4me1. Interestingly, RNA polymerase II most strongly correlates with marks close to the TSS and enhancers, which demonstrates that most polymerase is in either stalled states[254] or involved in the production of enhancer activity related transcripts (eRNAs).[256] Far less correlation exists between RNA PolII and H3K36me3, which is consistent with H3K36me3 being deposited during transcriptional elongation and therefore enriched towards gene 3' ends. Further, even though they locate to different parts of

genes H3K36me3 best correlates with H3K4me3, which is likely reflecting the cooperation of their writer enzymes in TrxG complexes.[66]

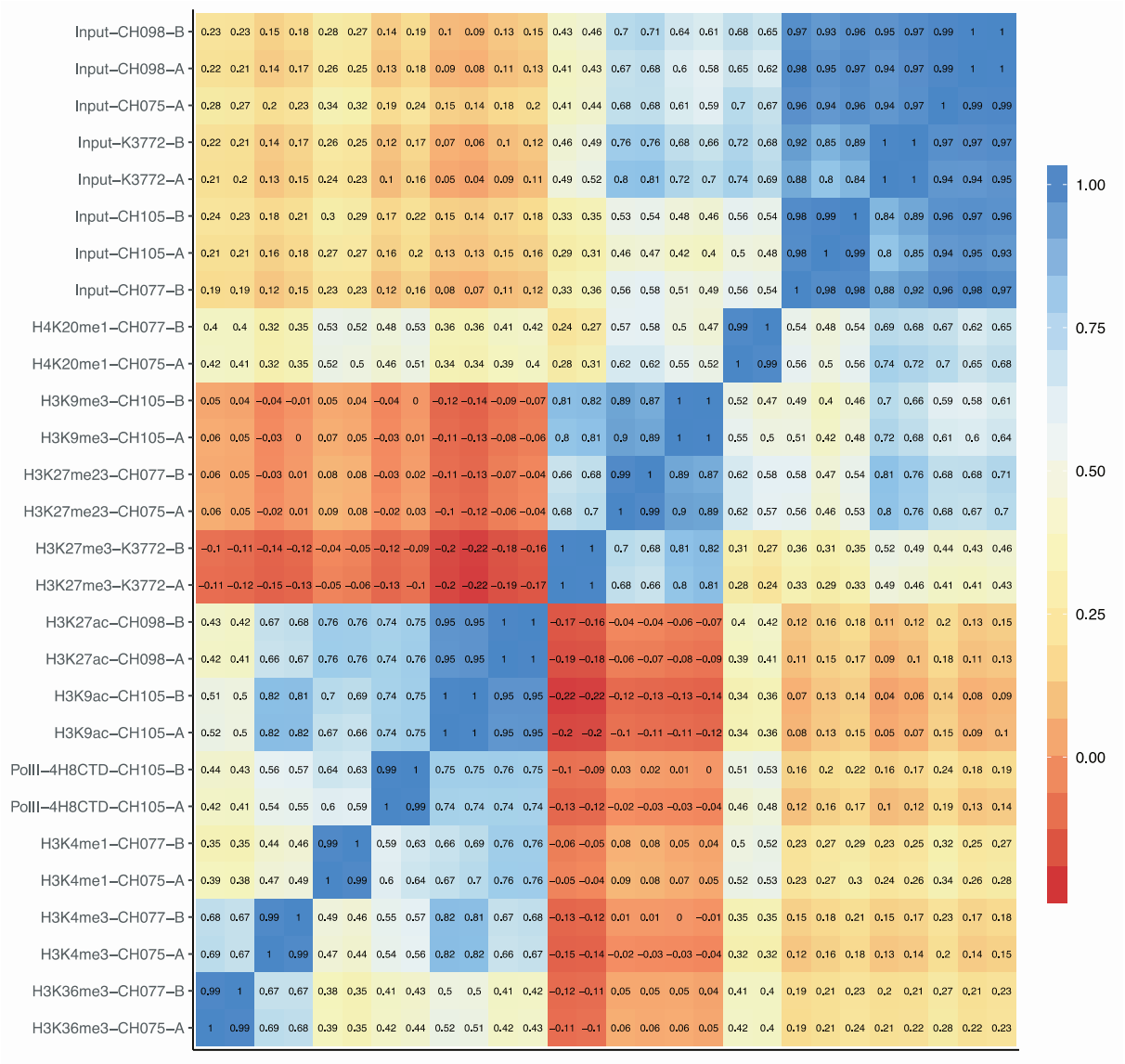

Figure 8: ChIP-seq samples correlate by replication and genomic proximity. Mapped reads from each individual ChIP-seq replicate were binned using 10kb windows. Pairwise Pearson correlation coefficients of the resulting read counts were calculated using Deeptools. Samples were sorted using hierarchical clustering and visualized as a heatmap. Colors indicate the correlation according to the scale on the right.

Underlining their assignment into active and repressive marks both classes of marks lack correlation and in part even anti-correlate. These small correlation coefficients seem to be similar across all different active ChIP targets, and the strongest anti-correlation to active marks is observed with the canonical gene repression mark H3K27me3. This unbiased anti-correlation of

repressive marks with active marks indicates, that repressive marks, unlike active marks, do not target individual parts of genes, but spread over entire silenced domains.

In summary, the high pairwise correlation of replicates demonstrates, that histone modifications were reproducibly enriched, and the degree of correlation between different marks well reflects their described genomic proximity.

As an alternative way to assess cross-sample variation and clustering observable in the ChIP-seq experiments, I performed a PCA analysis on the ChIPs by using 10 kb bins to quantify read counts and calculating the PCA rotations using the 1000 most variable bins across all samples. The loadings of the resulting first three principle components were plotted (Figure 9A). The replicates cluster well together, further supporting that the replication was successful. Similar to the results from Pearson correlation there are 3 major groups, with the active marks clustering towards the top left, the inputs in the bottom center and the repressive marks spreading towards the bottom right. The active gene marks separate from the input samples across principle component 1 (PC1), and this component explains 44% of all sample variance. It is, however, important to mention that I assayed a larger number of active marks and any common variance in these samples, therefore, has a greater impact on the overall data matrix than the smaller number of repressive marks. Consequently, a higher fraction of variance does not automatically imply a higher degree of dissimilarity between input and active marks than between input and repressive marks.

The clustering of the active marks reproduces some observations that were made from the Pearson correlation coefficient heatmap (Figure 8). The two acetylations H3K9ac and H3K27ac cluster closely together, furthest away from the input. RNA polymerase II and H3K4me3 cluster close to these, while H3K4me1 is slightly more distant from the acetylations. Interestingly, H3K36me3 in this analysis is closest to H3K4me1 and not to H3K4me3 both along PC1 and PC2, which is in contrast to the correlation-based analysis. H3K4me1 is, too, placed by a TrxG member[66], but normally locates to enhancers. As with the correlation analysis, H4K20me1 is much closer to input than any other active histone mark. Further, it is the only active histone mark that is distinguished from input almost exclusively by PC2 and not by PC1.

Along this PC2, 33.4% of all sample variance is explained, and this component distinguishes both active marks (to the left) and repressive marks (to the right). Strikingly, the repressive mark that is most distant from the input and thereby most distant from active marks along PC2 is H3K27me3. This reflects the instructive role of H3K27me3s in silencing[97] and the less broad, more localized enrichment observed for it when compared to the other silencing marks. H3K27me2/me3 is close to the input samples, perhaps reflecting its broad genome distribution, while H3K9me3 is located in between H3K27me2/me3 and H3K27me3.

The third dimension of the PCA explains very little of the variance (only 7%) and only the input samples vary strongly along it. Active marks do however also separate from repressive marks, which are along this dimension most distant from the input. Therefore, this component likely recapitulates the generally poorer recovery of repressed chromatin regions in the input (Figure 7).

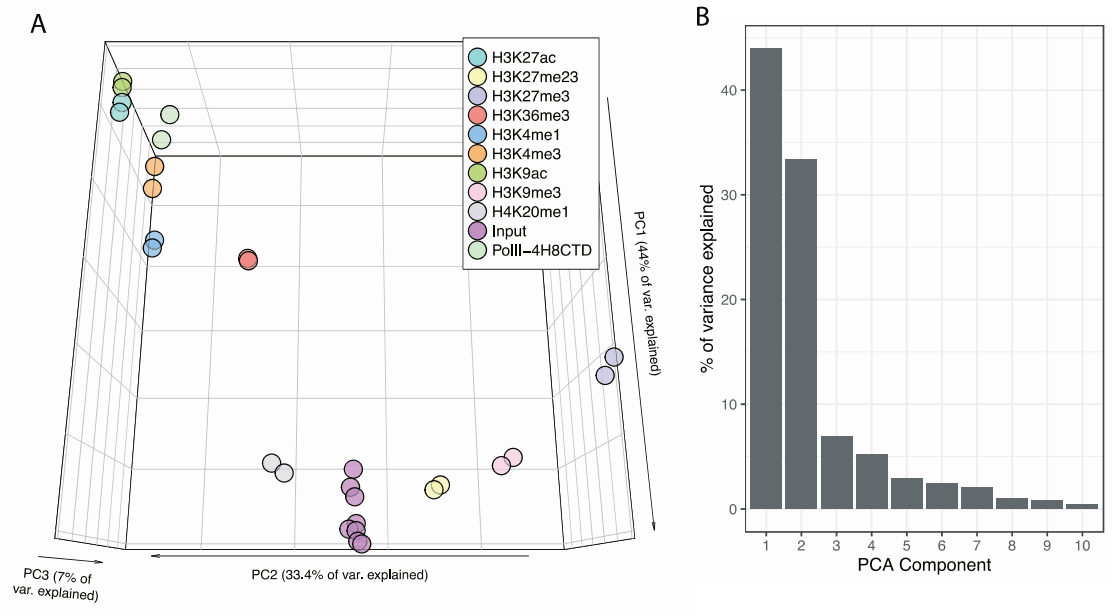

Figure 9: Principle component analysis of ChIP-seq samples separates active and repressive marks. Reads from all samples were counted using 10 kb genomic bins and the 1000 most variable bins were selected. The resulting counts were subjected to principal component analysis. A: Visualization of the first 3 components of the PCA with colors indicating the different antibody targets. B: Percent of total sample variance explained by the first 10 components.

The further PCA components encode even less of the total variance (Figure 9B) and no clear patterns that explain the separation along these dimensions were identified. Therefore, they were not plotted, limiting the PCA shown here to explaining 84.4% of all sample variance. In short, PCA well separates active and repressive histone marks, and the degree of separation reflects their mutual exclusiveness.

This PCA analysis comes with 2 caveats which I want to shortly address. First, the bins used for the PCA were selected for their high variance across all samples. While this is helpful to exclude excessive noise from the analysis, it might also limit the analysis to the more extreme bins, like regions around Hox genes or ribosomal genes. Second, the large size of the bins (10 kb), while also facilitating a stable, noise-free signal, may cause marks with a rather sharp distribution like

H3K4me3, which locates exclusively to TSS, to be captured poorly by this approach. This second point further extends to the correlation heatmap (Figure 8).

To quantify genome-wide localization of histone modifications on genes I therefore determined averaged signal intensity of these modifications along genes. Hence, I calculated $\log_2$ ratios of each ChIP-target over input in 25 bp genomic bins, normalized to the library size.

To be able to compare genes of different lengths I scaled the gene length of all genes to 10kb from transcriptional start site (TSS) to transcriptional end site (TES) and adjusted the signal tracks accordingly. Regions before the TSS and after the TES were added unscaled. The resulting gene-level profiles were separated into 2 groups using k-means clustering and the average signal intensity along the genes in each cluster was plotted (Figure 10A). Here, k-means clustering produces 2 groups that appear to be split into the genes with more repressive marks (in red and orange) and more active marks (in purple and black). Therefore, I labeled the 2 cluster groups as active genes and repressed genes.

Among the active genes, one striking feature is the depletion of all histone mark signals around the TSS. Instead, a strong peak of RNA PolII signaling is present in this position, which corresponds to a loss of nucleosomes in these positions.[257] The RNA polymerase II signal then trails out towards the TES. H3K36me3, on the other hand, is steadily increasing downstream of the TSS and only reaches its maximum slightly before the TES, consistent with prior observations.[20]

The repressive marks are lowest around the TSS of active genes and remain low throughout the gene body. On inactive genes, the repressive marks are much more prominent than active marks and show very little variation across the gene body, with only a slight enrichment towards the middle of the gene. Here, depletion of signal around the TSS is almost absent, perhaps only deriving from a minor subpopulation of sporadically transcribed genes. H3K27me3 is the most strongly enriched repressive mark on inactive genes. This may recapitulate its characterized role in gene repression, but an effect of antibody strength might also explain part of the difference.

Subsequently, I addressed the distribution of marks around the TSS, which is a hotspot for several highly localized marks. Here, a scaling of genomic regions was not necessary. After selecting appropriate histone PTMs to investigate, the data was otherwise processed as described above, using the same k-means method to separate active and inactive genes.

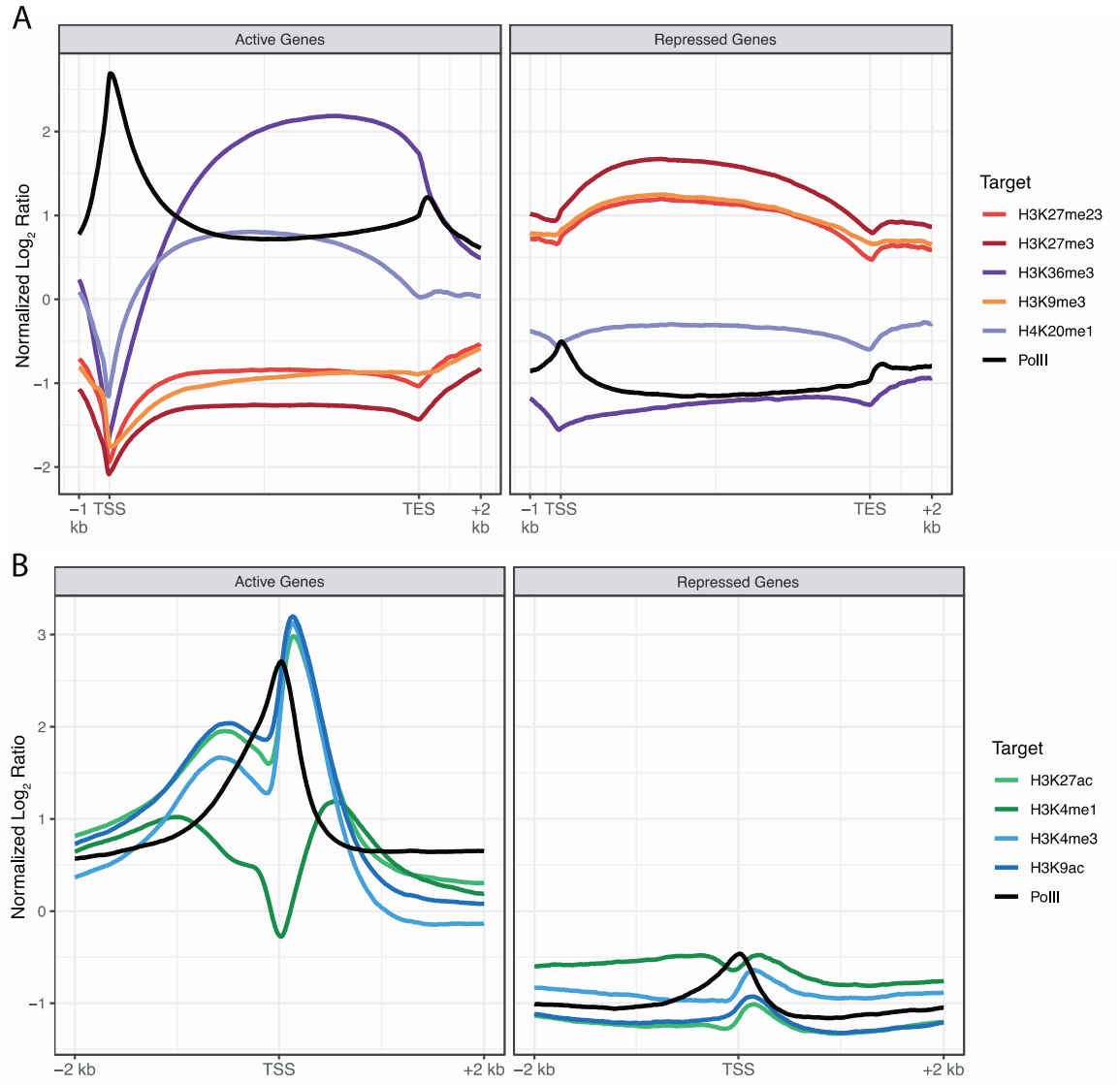

Figure 10: Histone marks distribute along genes concordantly with previous reports. Log_2 ratios of 25 bp coverage files for each ChIP target were calculated as ChIP-seq over input, both normalized to the total read count in that library. The resulting coverage files were quantified using Deeptools. A: A selection of marks was quantified by scaling all gene regions from the transcriptional start site (TSS) to the transcriptional end site (TES) to 10 kb and determining coverage. B: A selection of marks was quantified in a fixed window around the TSS. A and B: Genes were then split into 2 groups using k-means clustering and profiles were plotted.

Consistent with depletion of nucleosomes at the TSS I observe a lower signal of histone marks directly at the TSS of active genes. On active genes the RNA polymerase II signal is centered directly at the TSS, given the resolution of the ChIP, this likely reflects the previously described stalling of polymerase.[254] H3K27ac, H3K9ac, and H3K4me3 signals peak directly after the TSS with a shoulder towards the 3' region. H3K4me3 is most strongly localized towards the region directly downstream of the TSS, while the peaks for both acetylations are broader and also more

prominent upstream of the TSS. This reflects well the described distributions of these marks, with H3K4me3 enriched just around the TSS, while both H3K9ac and H3K27ac are present both at the TSS as well as at active enhancers and therefore more distant from the TSS.[55] H3K4me1 on the other hand is less enriched around the TSS both with regard to the sharpness of the distribution peak as well as the overall enrichment over inactive genes. This pattern fits well with the observation that that H3K4me1 is predominantly found at enhancers, which are not located at any fixed point relative to the TSS. In general, all these active marks are strongly depleted in inactive genes and are spread more evenly across the genomic region than in active genes. Together, the histone modifications show both a distribution across individual genes as well as a concordant separation along gene groups which well reflects established histone modification descriptions.[55]

In summary, I was able to establish a ChIP-seq protocol from primary plasmatocytes. This is to my knowledge the first ChIP from primary hemocytes and the first ChIP from any major *Drosophila* immune tissue. Further, the observed distributions of histone modifications are highly consistent with those that have previously been described for *Drosophila*, both with regard to their preference for active or repressed genes and to their localization on specific genomic elements, like TSS, enhancers or gene bodies. Interestingly, among the datasets which I generated H3K27me3 was most discordant with active gene marks, being both most enriched on inactive genes and most anti-correlated with active marks.

5.2.2 Multidimensional binomial expectation-maximization models

The strong anti-correlation to active marks raised the question of whether the H3K27me3 enrichment is based on only a small number of genes, such as Hox genes which are strongly targeted by H3K27me3 and permanently repressed, or if inactive genes, in general, are targeted by H3K27me3. Therefore, I developed an approach to quantify and categorize the enrichment of H3K27me3 (and histone modifications in general) on genomic regions or genes. I especially addressed whether I could detect one or multiple states of H3K27me3 enriched chromatin and how this explains the co-variance with gene state: There might be either a small set of highly H3K27me3 positive genes, or a larger set of intermediately enriched, or both.

Many algorithms have been proposed previously to detect such ChIP enriched regions, each with slightly different underlying mathematical models and their own advantages and disadvantages.[258,259] Perhaps best known and most used is Model-based Analysis of ChIP-Seq (MACS)[260], which identifies regions of interest by characteristic peak shifts between strands and estimates enrichment by ratios of sequencing depths using Poisson models. Many further

algorithms based on such Poisson models have been developed, including SPP[261] and SICER[262]. Others have proposed different models, like the bivariate HMM models implemented in histoneHMM[263].

One interesting approach is implemented in normR, which uses expectation-maximization (EM) based fitting of binomial models to identify enriched regions.[264,265] This implementation worked very robustly in my hands and has some advantages that the authors discuss: By modeling both foreground (enriched or signal) and background (non-enriched or noise) regions simultaneously, high sensitivities can be reached even with weak enrichment. Further, the count-based modeling reflects very well the additional uncertainty for shallow coverages of genomic regions. Last, the underlying model allows simultaneous fitting of more than 2 models to reflect chromatin states with different levels of enrichment in a ChIP-target or to determine differential enrichment between 2 ChIP-seq experiments. There are however additional interesting features that I noticed, which are not implemented in the normR package, but which are, in principle, supported by the underlying mathematics: First, unlike the normR implementation, genomic regions analyzed for enrichment do not have to be of fixed width but are flexible. While normR suggests 500 bp bins, the binomial model can also be applied to gene-level data. This is justified as long as the tracks for input and ChIP are similarly smooth across the whole region as they are across small bins, which is to say that the true ChIP signal is not localized to small peaks within the larger region of interest and therefore much of the region is in fact actually be background. Such uniform coverage is highly likely for H3K27me3 (also see Figure 10) and may, in fact, be valid for further histone marks. The second interesting property of binomial EM is that the model which is implemented in normR as a single dimension model or, in other words, requires exactly 1 ChIP and 1 input data set, can be extended to more dimensions. Such a multidimensional binomial expectation-maximization algorithm can be used to fit background and foreground components simultaneous across two or more replicates of a ChIP. In this case, each dimension or pair of ChIP and input is connected to the other replicates by the assignment to a component like background or foreground, while the parameters (or binomial probabilities) of the distributions in that dimensions may differ between different ChIPs. This approach is interesting, because ChIP replicates may produce different enrichments of signal to noise simply because of technical differences resulting for example from differences in the washing during the Immuno-precipitation. These differences would be measurable, but not biologically relevant. Therefore, building a model that captures this information is interesting to improve on the classic approach to handle replicates, which is to merge samples for peak calling (which might bias the results to

more strongly consider 1 replicate over the other) or to call peaks separately and identify overlapping enriched regions (which might strongly underestimate the likelihood of enrichment). I, therefore, decided to rebuild the binomial expectation maximization (EM) algorithm implemented in normR from the ground up to implement the possibility to hand over both, manually selected bins and to use any number of replicates when desired. In order to demonstrate the modifications made to the EM algorithm here, I first want to recapitulate the general math for the one dimensional (one sample pair) case. Figure 11 demonstrates the binomial model in the simplest case: Just one pair of input and ChIP, to which exactly one background (noise) and one enriched (signal) component should be fitted. In this case, after splitting the genome into bins using the desired method (fixed width, genes, etc.) the reads from ChIP and Input are counted in that region and throw them in an "urn" to draw from. Such a drawing process follows a binomial distribution:

$$\mathrm{p}(s_i, r_i|\theta_j) = \binom{s_i + r_i}{s_i} {\theta_j}^{s_i}(1-\theta_j)^{r_i} \quad (1)$$

With $i = 1, \dots, N$ the number of bins chosen, s_i and r_i the number of reads in the i-th ChIP or input bin respectively, and θ_j the probability of "drawing" a ChIP read. In the simple two-component model j can be 1 (background) or 2 (enriched) and θ_j is the same for every bin in the genome which belongs to component j. Here, therefore exactly 2 types of urns exist to draw reads from. However, the information whether a genomic region or bin is truly enriched is hidden and figuring out the assignment to the background or enriched "urn" is exactly the purpose of ChIP-seq peak calling.

Therefore, we must find a way to determine what is a good combination of estimates for both the binomial probabilities θ_j and the assignments of the bins. For ease of notation, I will call the true (unknown) assignment of the i-th bin to a component $z_{i,j}$, which is 1 if the i-th bin is in component j and else 0. Then the likelihood function can be derived from the binomial probability (1)

$$\mathcal{L}(s_i, r_i|\theta_j, z_{i,j}, \pi_j) = \prod_i^N \prod_j^m \left[\binom{s_i + r_i}{s_i} {\theta_j}^{s_i}(1-\theta_j)^{r_i}\pi_j \right]^{z_{i,j}} \quad (2)$$

where π_j is the fraction of bins belonging to component j and $j = 1, \dots, m$ the number of components (or different urns) used. Now a model can be found that contains the unknown parameters θ_j, $z_{i,j}$ and π_j such that it maximizes the likelihood function (2). This approach is also called maximum likelihood estimation and is useful, because it finds the best possible model,

given the mathematical assumptions, that explains the observed data. However, given the large number of unobserved variables $z_{i,j}$ the likelihood function cannot be maximized directly.

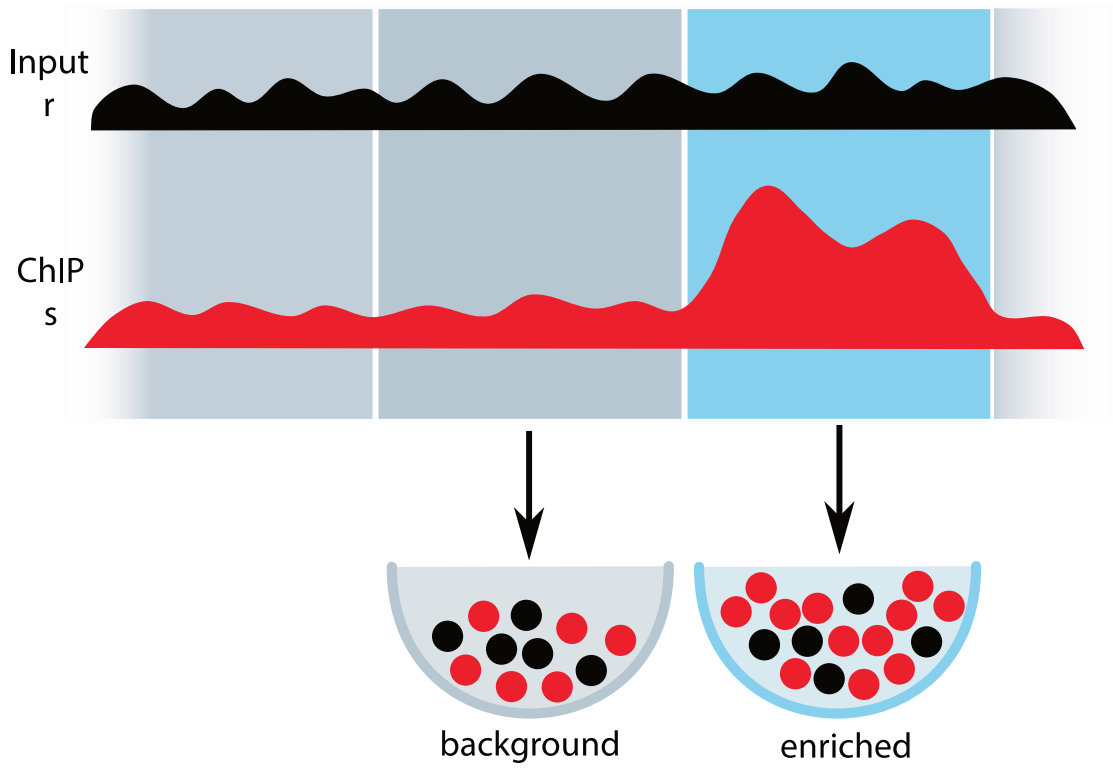

Figure 11: An urn model can be used to represent ChIP-seq enrichment. General representation of how the binomial model works for ChIP-seq. The reads are counted in bins (separated by white lines) for both the Input track (black, r) and the ChIP track (red, s). The resulting read counts of both samples can be considered as resulting from draws from an urn. Enriched regions have a higher ratio of red "balls" (reads) to black "balls" (reads) than background regions, which results in a higher probability of drawing ChIP reads from the enriched region "urn".

This is where the EM approach steps in. First, the likelihood function is log-transformed (2). Because the logarithm is a monotonically increasing function, any maximum on the log-likelihood function is also a maximum of the likelihood function. For the log-likelihood, we find the simpler equation

$$\ell(s_i, r_i | \theta_j, z_{i,j}, \pi_j) = \log \prod_i^N \prod_j^m \left[\binom{s_i + r_i}{s_i} {\theta_j}^{s_i} (1-\theta_j)^{r_i} \pi_j \right]^{z_{i,j}}$$
$$= \sum_i^N \sum_j^m z_{i,j} \left[\log \binom{s_i + r_i}{s_i} + s_i \log \theta_j + r_i \log(1-\theta_j) + \log \pi_j \right] \quad (3)$$

For the expectation-maximization algorithm $z_{i,j}$ is replaced by the expected value for $z_{i,j}$ $E[z_{i,j}|s_i, r_i, \theta_j, \pi_j] = E_{i,j}$, which results in the expected value of the likelihood function given the current set of parameters:

$$\mathrm{J}(\vartheta|\vartheta^{(t)}) = \sum_z q(z|x,\theta)\log p(x,z|\theta) =$$

$$= \sum_i^N \sum_j^m E_{i,j}\left[log\left(\binom{s_i + r_i}{s_i}\right) + s_i\, log(\theta_j) + r_i\, log((1-\theta_j)) + log(\pi_j)\right] \quad (4)$$

Now the EM algorithm optimizes the likelihood as follows: The model is initialized with a random selection of θ_j and π_j from the space of possible parameters. Then, the function is iteratively optimized by performing

E-step: $E_{i,j} = p(z_{i,j}|s_i, r_i, \theta_j, \pi_j)$ for each i

M-step: Set $\theta_j = \arg\max_{\theta_j} \mathrm{J}(\vartheta|\vartheta^{(t)})$ for each j

This process is continued iteratively until the change in the expected value of the likelihood function (4) drops below a chosen threshold. As such the EM algorithm can be considered a gradient ascend in which we iteratively maximize J with respect to $E_{i,j}$ and to the parameters θ_j and π_j. Others have demonstrated that such iterative accent at least converges to a local stationary point in J, such a point can however also be a saddle point or, rarely, a local minimum.[266,267] Most often, however, for well-behaved likelihood functions this will find at least a local maximum.[267] In order to maximize the chances of finding a global maximum, the EM-algorithm can impute a set of estimates several times from a different set of randomly chosen starting parameters. When these converged to their local maxima, the results which had the highest likelihood can be chosen ((2) or (3)) as the best model to explain the data.

Therefore, in the t-th iteration of the algorithm, we calculate the E-step with fixed θ_j and π_j:

$$E_{i,j}^{(t)} = \frac{{\theta_j}^{s_i}(1-\theta_j)^{r_i}\pi_j}{\sum_k^m {\theta_k}^{s_i}(1-\theta_k)^{r_i}\pi_k} = \frac{\alpha_{i,j}}{\sum_k^m \alpha_{i,k}} \quad (5)$$

and then maximize the parameter $\theta_k^{(t)}$ while keeping the $E_{i,j}$ fixed in the M-step:

$$\frac{\partial \mathrm{J}(\vartheta|\vartheta^{(t)})}{\partial \theta_k} = \sum_i^N E_{i,k}\left[\frac{s_i}{\theta_k} + \frac{r_i}{(1-\theta_k)}\right] \quad (6)$$

Which solves to

$$\theta_k^{(t)} = \frac{\sum_i^N E_{i,j} s_i}{\sum_i^N E_{i,j}(s_i + r_i)} \tag{7}$$

In a similar fashion, we can optimize for π_j:

$$\pi_k^{(t)} = \frac{\sum_i^N E_{i,j}}{\sum_i^N \sum_j^m E_{i,j}} = \frac{\sum_i^N E_{i,j}}{N} \tag{8}$$

This is the algorithm as it is implemented in normR.[264] When reimplemented in this fashion, it can address the first point made above: By applying it to non-fixed-width genomic intervals enriched genes can directly be identified, always under the assumption that the resulting ratio of ChIP and input reads still follow a binomial distribution. But the second objective has not yet been addressed: I want to transform the likelihood function in such a way, that multiple pairs of ChIP and input can be used for model building at once, ideally without being computationally too intensive.

Previously in (1), natural numbers were used in the binomial probabilities for the number of reads from ChIP s and the number of reads from input r. Now in the modified binomial probability function for the larger number of ChIP input pairs that there are d of such pairs (I will call them d dimensions). For the i-th genomic bin the numbers s_i and r_i can then be transformed into vectors $\boldsymbol{s}_i$ and $\boldsymbol{r}_i$ such that $\boldsymbol{s}_i, \boldsymbol{r}_i \in \mathbb{N}^d$ and the probabilities θ_j into vectors $\boldsymbol{\theta}_j \in \mathbb{R}^d$. We can then find:

$$\mathrm{p}(\boldsymbol{s}_i, \boldsymbol{r}_i | \boldsymbol{\theta}_j) = \prod_l^d \binom{s_{i,l} + r_{i,l}}{s_{i,l}} {\theta_{j,l}}^{s_{i,l}} (1 - \theta_{j,l})^{r_{i,l}} \tag{9}$$

with $l = 1, \ldots, d$ being the individual dimensions or pairs of samples. Such a probability function is still well normalized since for any vector $\mathbf{c} \in \mathbb{N}^d$ it is true that for the sum over all tuples of $\boldsymbol{s}, \boldsymbol{r}$ such that $\boldsymbol{s} + \boldsymbol{r} = \boldsymbol{c}$

$$\sum_{(\boldsymbol{s},\boldsymbol{r}):\, \boldsymbol{s}+\boldsymbol{r}=\boldsymbol{c}} \mathrm{p}(\boldsymbol{s}, \boldsymbol{r} | \boldsymbol{\theta}) = 1 \tag{10}$$

Therefore, the earlier likelihood function (2) can be modified by accounting for the multi-dimensional case:

$$\mathcal{L}(\boldsymbol{s}_i, \boldsymbol{r}_i | \boldsymbol{\theta}_j, z_j, \pi_j) = \prod_i^N \prod_j^m \prod_l^d \left[\binom{s_{i,l} + r_{i,l}}{s_{i,l}} {\theta_{j,l}}^{s_{i,l}} (1 - \theta_{j,l})^{r_{i,l}} \pi_j \right]^{z_{i,j}} \tag{11}$$

One could apply the brute force approach of finding solutions to every parameter by adding the additional index l, but this makes computation inefficient for higher-dimensional cases. Here, some matrix notation can help save a lot of time. By using the representations:

$$\boldsymbol{R},\ \boldsymbol{S}\ \in\ \mathbb{N}^{Nxd} = \begin{bmatrix} s_{1,1} & \cdots & s_{1,d} \\ \vdots & \ddots & \vdots \\ s_{N,1} & \cdots & s_{N,d} \end{bmatrix},\ \boldsymbol{\Theta} \in\ \mathbb{R}^{dxm} = \begin{bmatrix} \theta_{1,1} & \cdots & \theta_{1,m} \\ \vdots & \ddots & \vdots \\ \theta_{d,1} & \cdots & \theta_{d,m} \end{bmatrix},$$

$$\mathbf{E} \in\ \mathbb{R}^{Nxm} = \begin{bmatrix} E_{1,1} & \cdots & E_{1,m} \\ \vdots & \ddots & \vdots \\ E_{N,1} & \cdots & E_{N,m} \end{bmatrix},\quad \boldsymbol{\Pi} \in\ \mathbb{R}^{Nxm} = \begin{bmatrix} \pi_1 & \cdots & \pi_m \\ \vdots & \ddots & \vdots \\ \pi_1 & \cdots & \pi_m \end{bmatrix}$$

we can find the following simplifications. For the E-step we use

$$\log \mathbf{A} = (\log \alpha_{i,j}) = \boldsymbol{S} \times \log \boldsymbol{\Theta} + \boldsymbol{R} \times \log(\mathbf{J} - \boldsymbol{\Theta}) + \log \boldsymbol{\Pi} \tag{12}$$

with the $\mathbf{J}$ being the matrix of ones. From there we adapt from (5):

$$E_{i,j}^{(t)} = \frac{\alpha_{i,j}}{\sum_k^m \alpha_{i,k}}$$

We then turn to the M-step. The calculation of π_j is described in (8) and requires no modifications, since the number of dimensions d does not apply to this parameter. For the binomial probabilities $\theta_{j,l}$ we calculate

$$\log \boldsymbol{\Theta} = (\log \theta_{j,l}) = \log(\boldsymbol{S} \times \boldsymbol{E}^T) - \log\big((\boldsymbol{S} + \boldsymbol{R}) \times \boldsymbol{E}^T\big) \tag{13}$$

Using formulas (12) and (5) for the E-step and formulas (8) and (13) for the M-step, the EM-algorithm can then be applied to find the maximum likelihood estimators (MLEs) of the binomial probabilities $\theta_{j,l}$ as well as the probabilities $E_{i,j}$ with which each genomic region i belongs to a component j. Consequently, I implemented this multidimensional binomial EM as an R function (see source code in Code Snippet 1). It can be used by calling the function BinomEMwrapperParallel from EMcalc.R using the parameters N, which is the matrix of the total number of reads in each bin $\boldsymbol{N} = \boldsymbol{R} + \boldsymbol{S} \in \mathbb{N}^{Nxd}$, S, which is the number of reads in the ChIP sample $\boldsymbol{S} \in \mathbb{N}^{Nxd}$, and k, which is the number of components m which I will later also call clusters or states. Further, the number of attempted imputations can be specified (it defaults to 10), that is, how many random starting parameters should be tried for the optimization. I found that most starting parameters will arrive at the same MLEs, but on highly complex samples that could contain a larger number of local maxima, it might be useful to increase this further. Additionally, the tolerance can be specified, which will determine at what change of likelihood between iteration steps the algorithm will consider the estimates to have converged. A maximum number of iterations was also implemented, which serves as an escape, should the

algorithm ever diverge, though I have never observed this. Last, to increase speed, the function was made multithreading capable. The multithreading, in this case, is simply implemented by running each imputation as a separate thread, which means that for ideal performance the number of threads should be adjusted to the number of imputations to be tried.

Each imputation starts by randomly sampling $\boldsymbol{\Theta} = \theta_{j,l} \in [0,1]$ and $\pi_j \in [0,1]$. Then the iterative improvement of the parameters starts: First, in the E-step, the expectations are calculated according to formulas (12) and (5), then in the M-step the maximum likelihood estimator of the parameters $\theta_{j,l}$ and π_j are calculated according to (8) and (13). Now the improvement in the log likelihood (see (11)) needs to be calculated to determine if the model has converged. Because the log-likelihood is essentially

$$\log \mathcal{L}\left(\boldsymbol{s}_i, \boldsymbol{r}_i \middle| \boldsymbol{\theta}_j, z_j, \pi_j\right) = \sum_{i,j} \log \alpha_{i,j} + d \sum_{i,l} \log \binom{s_{i,l} + r_{i,l}}{s_{i,l}} \tag{14}$$

the difference in log-likelihood between iterations t and t+1 are easily found by

$$\Delta^{t,t+1} \log \mathcal{L}\left(\boldsymbol{s}_i, \boldsymbol{r}_i \middle| \boldsymbol{\theta}_j, z_j, \pi_j\right) = \sum_{i,j} \log {\alpha_{i,j}}^{t+1} - \sum_{i,j} \log {\alpha_{i,j}}^{t} \tag{15}$$

which can be derived from (12). Once that difference is smaller than the set threshold the model is considered converged and saved.

After all imputations are done all resulting models are compared for their log-likelihood and the model with the largest likelihood is chosen as the best model to explain the data. Last, each of the bins on which ChIP and input reads were counted is assigned to one of the components *j* by choosing that component with the highest expectation in the final best model. The function then returns this assignment, along with all other model parameters like the final parameters $\theta_{j,l}$ and π_j, and the expectations $E_{i,j}$.

The EM algorithm described here, therefore, improves on the algorithm implemented in the normR[264] package in several ways. First, I can apply it to any count data from genomic bins, both using a selection of genomic bins and using bins of different widths. This would have been possible with the normR package too but would have required the calling of private functions that are not meant to be called by the user. More importantly, however, the model is expanded to work with any number of ChIP and input pairs. This can, for example, be used to more accurately call p-values when using ChIPs with replicates. If θ_B is the binomial probability of a background (or noise) component in the normR[264] package, they suggest to calculate the p-value for the null hypothesis that the bin i derives from the background component as

$$p = \sum_{z > s_i}^{s_i + r_i} \binom{\mathrm{z} + r_i}{\mathrm{z}} {\theta_B}^{\mathrm{z}} (1 - \theta_B)^{r_i} \tag{16}$$

Helmuth, et al. [264] suggest to handle replicates by simply summing the counts for input and ChIP, but this will lose a lot of information contained in the replication and therefore misrepresent the p-value. In contrast, with my implementation the p-value can instead be calculated by fitting multidimensional models and finding:

$$p = \sum_{z_l > s_{i,l}}^{s_i + r_i} \prod_{l}^{d} \binom{s_{i,l} + r_{i,l}}{s_{i,l}} {\theta_{j,l}}^{s_{i,l}} (1 - \theta_{j,l})^{r_{i,l}} \quad (17)$$

using the vector forms of $\boldsymbol{s_i}$ and $\boldsymbol{r_i}$ and summing over all $\boldsymbol{z}$ such that every $\boldsymbol{z_l}$ is larger than $\boldsymbol{s_{i,l}}$.

The EM model can however also be used for other purposes than calculating more accurate p-values: It can be used with a set of different ChIPs on different targets, thereby finding categories of bins with co-varying binomial probabilities.

In summary, I modified existing methods to fit binomial mixture models through an expectation-maximization algorithm, such that it can be used to handle replication. This algorithm can also be used to fit models to larger ChIP-seq data collections to predict principle chromatin states.

5.2.3 Histone PTM chromatin state models by EM

To verify the functionality of my newly implemented algorithm I applied it to the ChIP data I generated from *Drosophila* plasmatocytes. First, I showed that the algorithm can identify different types of H3K27me3 enrichment on genes. Because H3K27me3 is distributed almost evenly across genes both in the active and silenced gene subgroups (Figure 10) I can use each gene span as a region of interest and count the total number reads from both ChIP and input replicates overlapping with each of these gene regions. I can then use those counts in the EM model and call different components or clusters. Initially, I set the number of EM clusters to 3 since that well explained the distribution of signal shown in Figure 7, where there are 3 types: H3K27me3 low genes which are transcribed, H3K27me3 medium genes which are non-transcribed and H3K27me3 high state genes which are non-transcribed developmental genes. The EM algorithm then processed the number of ChIP and input reads and grouped the genes into 3 different components based on their probability to derive from the fitted binomial distributions.

To better visualize the results, I will first introduce one definition I will use throughout my thesis: I here define the enrichment for each gene as the expected value of the binomial probability for gene i:

$$Enrichment(gene_i) = \frac{s_i}{s_i + r_i} = \frac{\#\ reads\ in\ ChIP\ at\ gene_i}{\#\ reads\ in\ ChIP\ and\ input\ at\ gene_i} \quad (18)$$

where the number of reads is normalized to the total number of reads generated from the respective library. This definition of enrichment slightly differs from the usual definition of enrichment which is the log ration of ChIP over input, but has some advantages: First, it works well even when a library was only sequenced shallow, such that a number of bins have zero reads overlapping with them. When using the log-ratio, if either input or ChIP is zero over a bin, the log ratio cannot be calculated correctly for that bin. Therefore, when using log ratios, pseudo-counts often have to be added to prevent zero values. With my definition of enrichment, this is only a problem if both ChIP and input are zero (in which case the analysis of that bin likely makes no sense). Second, my definition of enrichment works very well with the concept of binomial models, because the value I find is also the maximum likelihood estimator for the binomial probability of that isolated genomic region. Last, the profiles generated by calculating enrichment as in (18) are overall similar and it will for any non-zero pair of input and ChIP reads produce the same rank of a bin as the log ration method.

$$\frac{s_1}{s_1 + r_1} > \frac{s_2}{s_2 + r_2} \Rightarrow \frac{s_1}{r_1} > \frac{s_2}{r_2} \tag{19}$$

Further, there are some interesting properties in this way of calculating the enrichment, as it is always between 0 and 1, and 0 will mean exactly no reads in the ChIP sample, while 1 will mean no reads in the input. Because of these advantages, I will use (18) throughout my thesis to calculate enrichment.

After applying the EM algorithm to the H3K27me3 data I plotted the enrichment of the two replicates of H3K27me3 ChIP-seq I performed from *Drosophila* plasmatocytes, normalized over the matching input sample (Figure 12). In the scatter plot, each dot represents a single gene with its corresponding enrichment in each replicate. Using this analysis, the two samples correlate very well on a gene level (r = 0.988, p-value < 2.2e-16), similar to what I observed previously when applying 10 kb bins for sample correlation (Figure 8).

Above the scatter plot there is a histogram of the binned frequencies that were observed in replicate A, and the histogram is, therefore, a projection of the data from the scatterplot. Each gene was colored according to the component it was assigned to by the EM algorithm, and I will refer to these components as states. This new EM-based method of clustering worked well, as is apparent by the distinct regions into which all genes of a state fall. From the histogram, it is clear that the 2 peaks that can be observed in the enrichment plot are each covered by a different state, where the low state (green) covers the genes with the lowest levels of H3K27me3, while the medium state (orange) coverers H3K27me3 intermediate genes. Further, the genes that are in-

between these peaks are split around the middle, such that genes below this (soft) threshold are assigned to the low state, while genes with higher enrichment belong to the medium state. The high state (purple) marks the genes with very high H3K27me3 levels, which represent only a minor population among all genes.

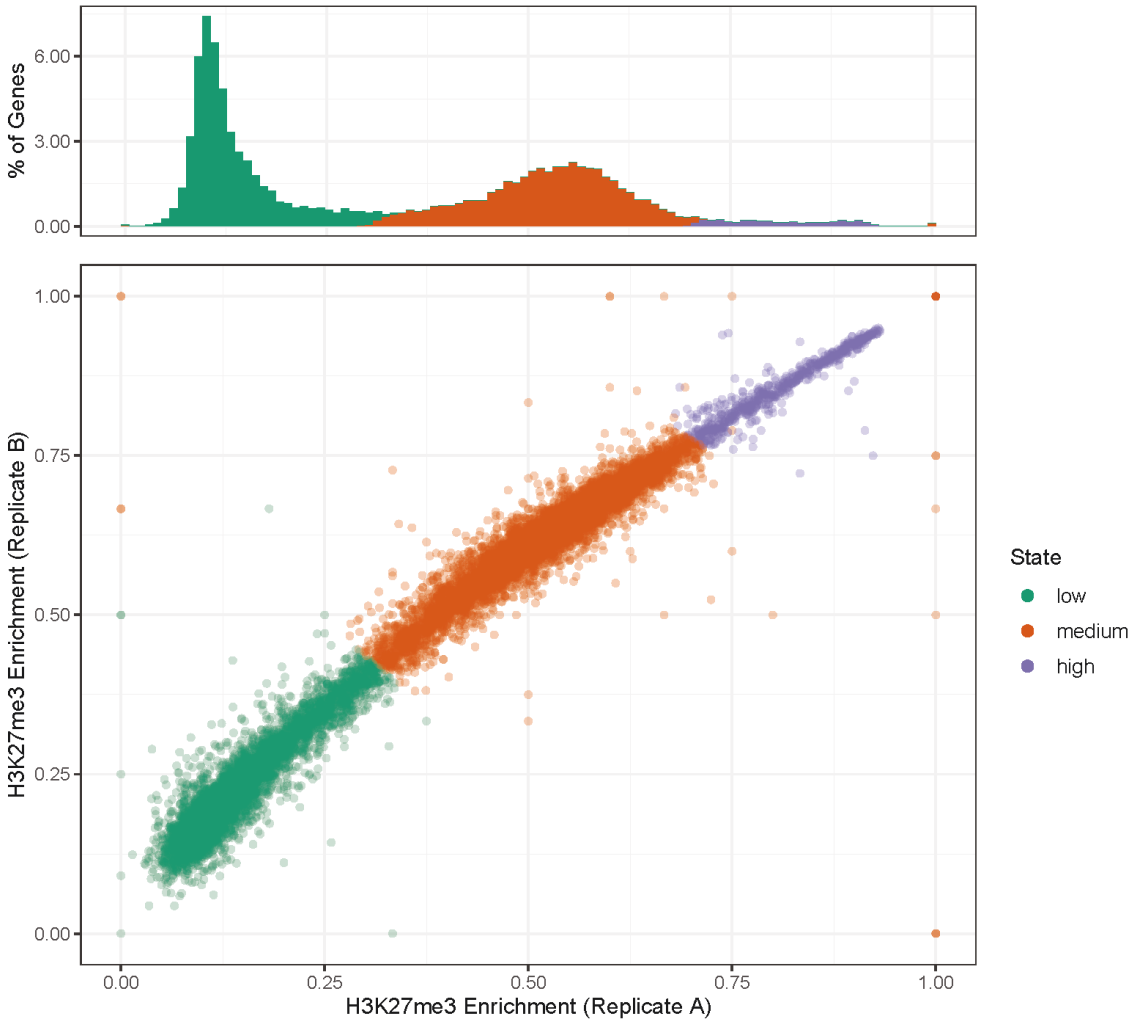

Figure 12: Gene level H3K27me3 EM-clustering produces three distinct gene states. Gene level read counts of H3K27me3 ChIP-seq samples and corresponding input samples were clustered into 3 groups using a binomial EM-clustering. Enrichment (ChIP / ChIP + Input) was calculated for each replicate ChIP input pair. Lower panel shows the correlation of this gene-level enrichment between the two replicates and each dot is a gene. Genes were colored by their EM-state. The top panel is a projection of replicate A into a histogram.

So, what does this state assignment mean in biological terms for the respective genes? The low state with its very low H3K27me3 signal is well explained by the absence of H3K27me3 from actively transcribed genes. The H3K27me3 high state predominantly contains developmental

genes, like the Hox genes and e.g. *eve*, that I used as an example in Figure 7. Therefore, the state seems likely to represent stably repressed genes, for which PcG proteins and strong PREs are instrumental.[76,89,95] The remaining H3K27me3 medium state apparently defines non-expressed genes which are not targets of strong PREs but represent an intermediate case. The question remains though, if the targeting of these genes by PcG proteins and the placement of H3K27me3 is functionally relevant at these genes, or if it is merely incidental. In short, EM clustering predicts a three-state model for H3K27me3 levels, which recapitulates the established Polycomb chromatin state and H3K27me3 low state, but also adds an H3K27me3 medium state, that is clearly distinct from the other two states.

To investigate whether a combination of histone modifications helps to infer more about the different H3K27me3 states, I applied the algorithm to the complete data set of plasmatocyte ChIP-seq runs. In this case, given the much more diverse distribution of the signals for the different targets (see Figure 8 and Figure 10) than in H3K27me3 alone, using gene wide signals is no longer an option. Therefore, I instead divided the genome into 200 bp bins and counted the reads overlapping with each of these bins for each pair of ChIP replicate and its matching input. This generates tracks for 20 ChIPs, 2 replicates of each ChIP of RNA-PolII, H3K4me1, H3K4me3, H3K9ac, H3K9me3, H3K27me3, H3K27me2/me3, H3K27ac, H3K36me3, and H4K20me1. Further, because not all ChIPs were done at the same time, I have 8 different input samples, each belonging to a different set of ChIPs. This genome-wide fixed bin width data was then applied to the EM algorithm in much the same way as the gene level H3K27me3 data was before. In contrast to the gene level H3K27me3 analysis however, EM analysis on the whole set of marks generates clusters that are characterized simultaneously by the enrichment levels of multiple marks. Therefore, a clear ChIP background cluster might not appear. This is due to the disjoint, or non-correlating presence, of histone marks: Where H3K27me3 is maximal, H3K36me3 is minimal (which is to say it is at its background level) and the other ways around. The clusters, therefore, are more likely to represent different chromatin states, each with a set of histone modifications, that are enriched over their background level, and others that are not enriched. Here, I tried several different component numbers, but settled on 7 clusters (later called states), because this number of states seemed to best explain the data. As others have mentioned[55,69], there is no clear way to determine the optimal number of states to fit, and any choice must be to some degree arbitrary.

First, I tested if the histone mark enrichments observed in the resulting 7 states argue for a chromatin state model. Consequently, I separated the bins according to their assignment by the EM algorithm and calculated the enrichment in each bin according to (18). I then formed the

mean of those enrichments from each bin first by the state to which they belong and then by the histone modification targeted in the immuno-precipitation. The resulting data were hierarchically clustered across columns and rows and plotted as a heatmap (Figure 13A). This heatmap shows a clear separation of the different histone marks. The left 3 states (states 3, 1 and 5) are depleted of active marks (which are towards the top of the heatmap) but have increased levels of H3K27me3, H3K27me2/me3 and H3K9me3. These states are therefore likely to represent repressed chromatin states. The right 3 states (states 6, 4 and 7) on the other hand are strongly depleted in all repressive marks, but they are all enriched in at least one of the active histone modifications. State 2 in the middle appears to represent more of a transitory type of chromatin, though the comparably higher levels of active marks seem to indicate some type of active chromatin.

By comparing each chromatin state to published data, the underlying biological terms can be identified. The rightmost state 7 is very strongly enriched in H3K4me3 and H3K9ac, but RNA polymerase II and H3K27ac are also high. It seems therefore likely that this state represents active TSS, where H3K4me3 enrichment has been well described.[268] In state 4 many similar marks are characteristic, but instead of H3K4me3, I find a strong enrichment of H3K4me1. This suggests that these regions represent active enhancers.[22] State 6 has a different pattern of enriched marks and here H3K36me3 is by far the most enriched, while RNA polymerase II is intermediately enriched. This is consistent with regions of transcript elongation on active genes.[20] At the same time state 6 also has somewhat increased levels of H3K4me1 and H3K27ac which indicates the presence of some enhancers in the regions covered by this state. The intermediate state 2 is not strongly enriched in any one mark. I do however observe only low levels of H3K9me3 and H3K27me3, while H3K4me1 and RNA polymerase II are still above background levels. Therefore, it seems most likely that these are weakly bound enhancer regions, where both H3K4me1[22] and RNA PolII are present.[256] The 3 repressive states cluster on the left. Among them, state 5 is strikingly high in H3K27me3 and H3K9me3, but not the highest in H3K27me2/me3. This likely indicates that in these regions all histone residues of H3K27 are fully methylated and no H3K27me2 is left. Such high levels of H3K27me3 would be expected from silenced developmental genes like *eve* (which I have shown in Figure 7) or the Hox genes. Therefore state 5 probably identifies those genomic regions which are most strongly targeted by PcG proteins.[43] State 1, on the other hand, is also high in H3K9me3, but much lower in H3K27me3 and instead high in H3K27me2/me3. This high levels of H3K9me3 without the strong histone H3 lysine 27 trimethylation indicates that this state represents structural heterochromatic regions like centromeres and telomeres or transposons.[269] Last, in state 3 all repressive marks are present,

with a slight enrichment of both H3K27me3 and H3K27me2/me3 over H3K9me3. These genomic state therefore likely corresponds to the H3K27me3 intermediate state, which I previously identified on a gene level (Figure 12). It is also interesting, that the lower levels of H3K27me3 are not complemented by increased levels of H3K27me2/me3, which is highest in state 1. Therefore, this state is not one of high H3K27me2 with intermittent H3K27me3, but instead a H3K27me3 positive chromatin state. Importantly, none of the active marks is any higher here than in the other 2 repressive state 1 and 5. This indicates that these chromatin regions do indeed represent repressive chromatin states and not mere intermediates between active and repressive states.

The resulting states also vary greatly in their size (Figure 13B). Most of the 200 bp genomic bins appear to fall into state 3, while other states like state 7 or state 5 have far fewer elements. Perhaps this corroborates earlier observations from Figure 7 and Figure 10: Some of the marks like H3K4me3 are highly localized to only a very small parts of the genome (in the case of H3K4me3 only to TSS of active genes) while other marks like H3K9me3 or H3K27me2/me3 are broadly spread and can be found in many regions, both on silenced genes, on intergenic regions and on structural heterochromatin.

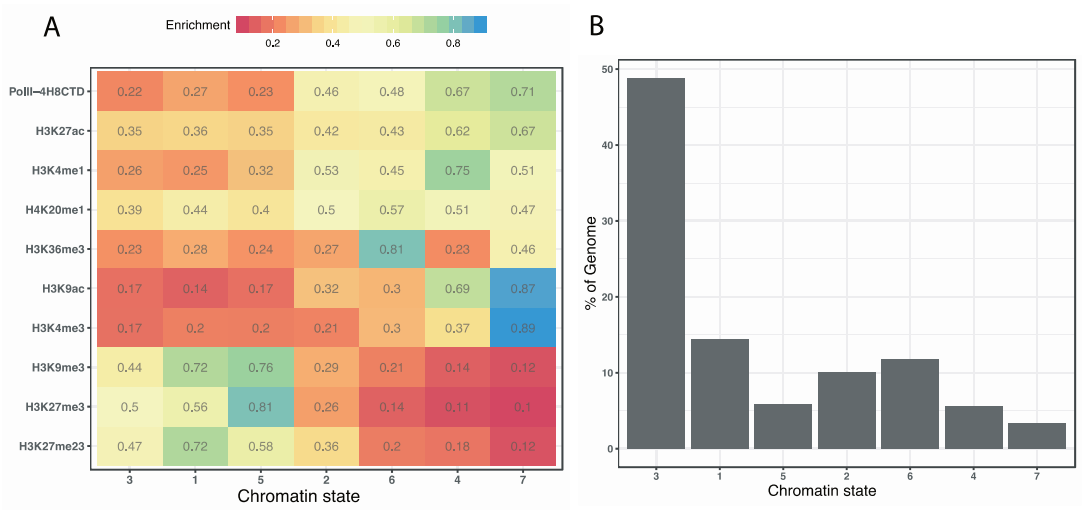

Figure 13: Genome-wide EM-clustering produces a seven-state chromatin model. Genome-wide read counts were quantified in 200 bp bins for ChIP and input samples. A binomial EM-model was used to cluster the resulting counts into 7 different components (states). A: Average enrichment of different ChIP signals in each state of the EM model, where rows correspond to ChIP targets and columns to EM states. Rows and columns are sorted by hierarchical clustering. B: The fraction of 200 bp bin in each component (state) representing the percentage of the genome bins belonging to each chromatin state.

Indeed, state 7, which is likely localized at TSS of active genes, accounts for the smallest fraction of the genome. The states for elongating transcription (state 6) and for enhancers (4 and 2) account for a larger part of the genome. State 1, which appears to be structural/constitutive

heterochromatin makes up only 14% of the genome, which is somewhat smaller than published data of total heterochromatin content in *Drosophila*.[270,271] It is important to consider that short-read sequencing approaches like Illumina sequencing are notoriously bad at covering these regions because of their repetitive nature, and the common reference assemblies are therefore missing much of the centromeric and telomeric regions.[272] Interestingly, of the two H3K27me3 targeted states, states 5 and 3, state 5 only covers a small fraction of the genome, which is consistent with it accounting for the relatively few key developmental genes strongly targeted by PcG proteins, while state 3 is by far the biggest of all states. Two factors might contribute to the spread of this state: First, I showed that the silencing marks are not localized exclusively to TSS, enhancers or gene body (Figure 7 and Figure 10) but spread across both genes and intergenic regions, and second, it appears that many genes fall in this intermediate category represented by state 3 (Figure 12).

In summary, the chromatin states defined here corroborate the established roles that marks have and correlate well with published models. A clear distinction is that this new EM model defines an additional repressed state where previous models proposed only a single H3K27me3 positive state.

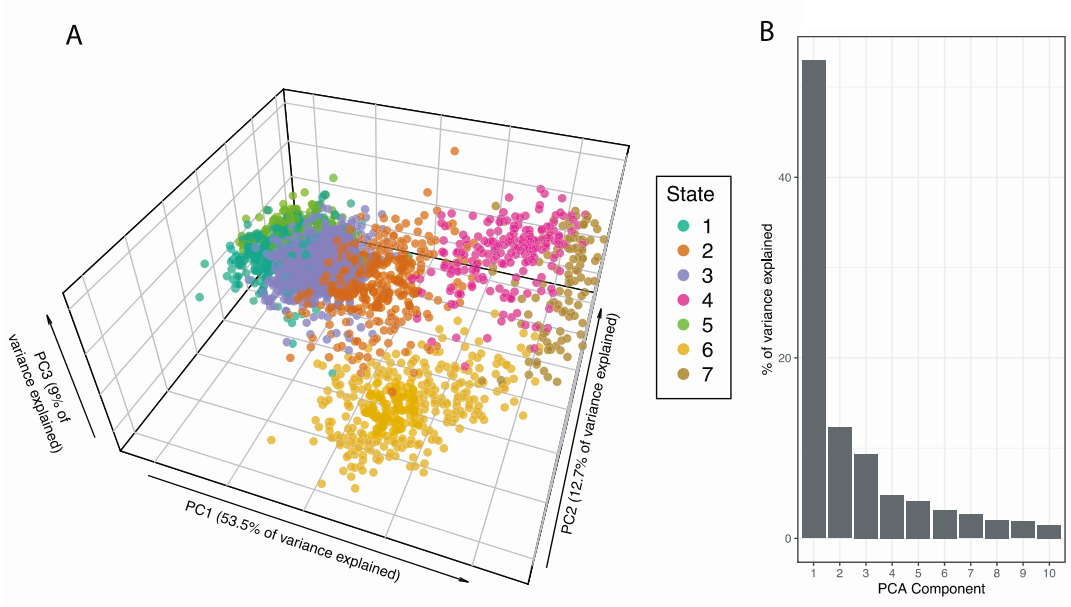

Figure 14: Chromatin states separate by their function in PCA. ChIP-seq signal was processed into bins as in Figure 13 and states were determined by EM. The same bins were then subjected to PCA. A: The enrichment in bin level read counts were subjected to PCA and from this analysis points were randomly sampled to plot. The first 3 dimensions of the PCA are plotted, where each point represents a bin. Bins are colored according to the EM-cluster into which they fall. B: Percentage of variance explained by each component of the PCA from A.

In order to test for the robustness of my chromatin state model, I tested whether PCA, would allow similar state separation. I first calculated the enrichment according to (18) from the signal that was detected in each ChIP across all 200 bp bins and which was also used to build the EM model. The resulting data matrix, in which each column represents a genomic bin and the enrichments in the individual ChIP replicates represents the rows or features, was then used to calculate the PCA rotations. From the rotation matrix, in order to prevent overcrowding the plot with 687743 data points, I randomly sampled 3000 bins for the final PCA plot (Figure 14A). The individual bins which represent dots in the PCA were then colored according to the chromatin state to which they were assigned during EM clustering. By the separation of the colors I can show that these mathematically different approaches of reducing data complexity are well concordant in their separation of genomic bins.

Along the principal component 1 (PC1), which by itself explains 53.5% of all the variance observed across the genomic bins, the active chromatin state (states 4, 6 and 7) to the right separate from the repressive states (states 1, 3 and 5) to the left. The state 2 which appeared intermediate in my earlier analysis is also in between the 2 groups here. Further, the states 4 and 7, which I proposed to be enhancers and TSS respectively, separate along the PC2 from the state 6, which are the H3K36me3 positive 3' end of active genes. Among the later principal components, I find that PC3 separates chromatin state 4 from the bulk of the other states. Interestingly chromatin state 2 slightly co-varies along PC3 with state 4, which is concordant with my previous hypothesis that state 2 represents a slightly less active type of enhancer chromatin. The last component that yields any separation of states is component 4, which slightly separates EM state 7 from the other state. Together these four components account for 79.4% of the variance (Figure 14B). Interestingly the repressive states 1, 3 and 5 do not separate well along either of the principal components and in the PCA the transition between their locations is fluent. It is not clear, whether this represents true biology (it is possible that these chromatin states are facets of a continuum), or if the large discrepancy in active marks (7 in total) and repressive marks (3 different histone modifications) are responsible for the poor performance of the PCA in the separation in repressive states. In summary, the PCA supports the EM-based chromatin state model by demonstrating that principal components can separate the predicted chromatin states in an unbiased fashion.

To characterize how genes associate with chromatin states I counted genes by their overlap with the 7 chromatin states from the EM model. Genes were classified as being in a given chromatin state if any of its exons had any overlap with a chromatin state. Using this definition I do expect

genes to overlap with several chromatin states: Given for example that chromatin state 7 should be at the TSS of active gene and state 6 at 3' ends of active genes, I expect most active genes contain both of these states at once. The overlap of genes with states is quantified in Figure 15.

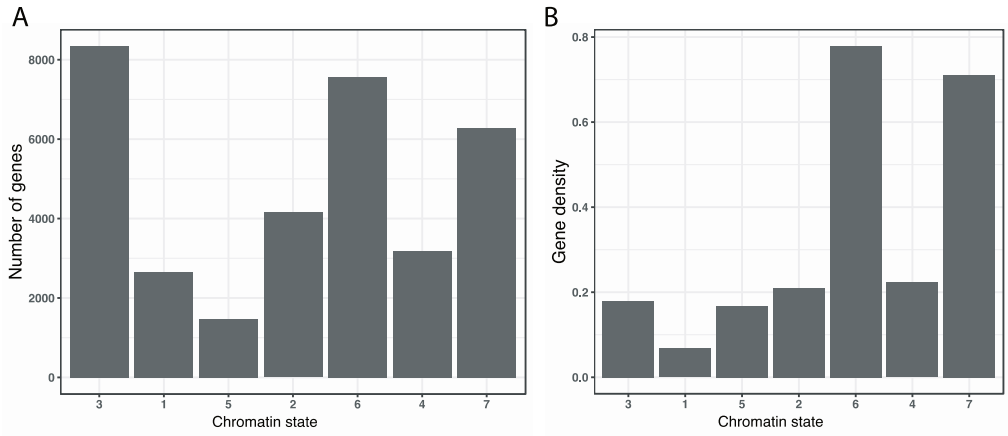

Figure 15: Chromatin states are distinct in their relative gene content. All genes from genome assembly dm6.91 were checked for overlap with chromatin state states of bins from the binomial EM model. Any overlap of 1 or more bp caused the gene to count towards overlapping with that state. A: Absolute number of genes overlapping with each state. B: Fraction of bins in a state that overlap with a gene.

This overlap of genes and states (Figure 15A) only partially reflects what may be expected based on previous chromatin state models. Chromatin states 6 and 7, which should overlap with the 3' exons and 1st exon of active genes respectively, are indeed associated with a large number of genes. At the same time, state 2 and 4, which appear to represent enhancers, also seem to overlap with exons, but a lower number of them. Compared to previous reports it also seems surprising, that state 1 should overlap with so many genes since it represents heterochromatin, which is supposed to be gene-poor.[273] In fact, here it appears that state 1 (heterochromatin) overlaps with more genes than state 5 (Polycomb chromatin), even though the latter's primary function is gene silencing. To solve these discrepancies, I calculated gene density, which I define as the fraction of all bins in a chromatin state that overlap with any exon by at least 1 bp (Figure 15B). Based on gene density the chromatin states fit much better with the published data. States 6 and 7 are very gene (exon) dense, while state 1 is by far the least gene dense of the chromatin states. Further, chromatin states 3 and 5 are similarly low in gene density, likely because repressive chromatin marks like H3K27me3 spreads also to intergenic regions, unlike active marks in chromatin states 6 and 7. In addition, it has to be taken into account that introns are not counted towards this gene density, which likely affects chromatin state 5: many Hox genes are very rich in long introns, a feature which is otherwise uncommon in *Drosophila*, in which most

genes only have short or no introns.[274] Chromatin state 3, which was not previously identified in chromatin state models, is of particular interest with regard to gene density. This state is low in gene density, but it also covers a large part of the genome (Figure 13). It, therefore, overlaps with more than 8000 genes. The results from RNA-seq of plasmatocytes showed that about 9300 genes were not detectable at a level useful for statistical inference, which means that the number of genes in the repressive state 3 is by no means excessive. Only a small fraction of the non-transcribed genes from plasmatocytes are known to be PRE based PcG targets[275,276], which should fall into state 5, and the remainder of silent genes must, therefore, fall into another repressive chromatin state. This also matches my observations on the H3K27me3 gene states, where many genes appear to be in the H3K27me3 intermediate category (Figure 12, medium state). Together, gene density across different chromatin states reflects observations for the prior description of the states. Importantly, the novel H3K27me3 intermediate chromatin state 3 has similar gene density to the H3K27me3 high chromatin state 5 but covers a far larger number of genes.

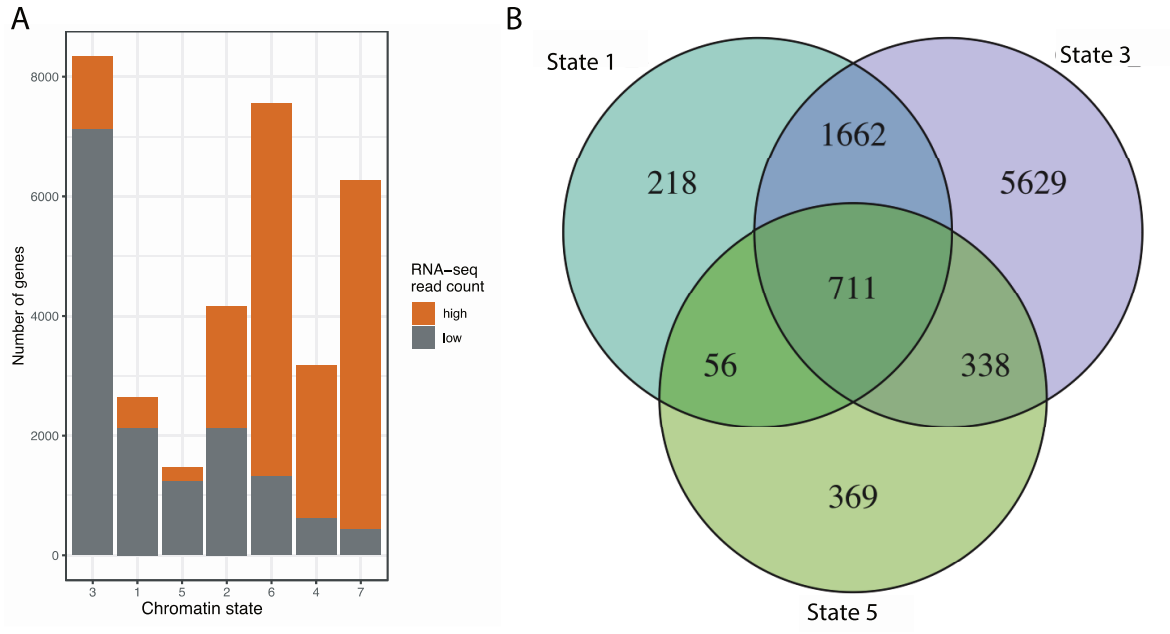

Figure 16: Repressive chromatin state overlap in a fraction of genes. A: Gene overlap of genes with chromatin state based on gene expression. Genes overlap was determined as in Figure 15A. Then RNA-seq expression levels were determined and any gene which is not at least covered by 1 read per million in any 4 samples in the RNA-seq is called RNA-seq low. B: Venn diagram of genes overlapping with either of the repressive states 1, 3 or 5. Genes overlaps are determined as in Figure 15.

Subsequently, I tested whether the overlap of a given gene with a distinct chromatin state is a good predictor of its transcriptional activity. Accordingly, I compared the EM clustering results to

the RNA-seq data from unchallenged plasmatocytes (Chapter 5.1.3). I, therefore, separated the genes scored in Figure 15A by their RNA seq read counts as, 'high' in RNA-seq counts if it had at least 1 read mapped to it per 1 million sequenced reads, and 'low' if it had less than that (Figure 16A). All 3 repressive states state 1, 3 and 5 overlap with a large number of low read count genes and only a small number of high read count genes. At the same time the active chromatin states 4, 6 and 7 show the reverse and overlap predominantly with high read count genes. State 2 is intermediate, overlapping to about the same extent with high and low read count genes. Additionally, the number of low read count genes, even when only considering the repressive states 1, 3 and 5, exceeds the total number of inactive genes of 9294. When summing up the number of genes from the RNA-seq low group from states 1, 3 and 5 I find that they add up to 10504 genes. However, as mentioned before, genes can overlap with several states. Therefore, when investigating the role of the individual chromatin states in gene repression, it is important to be aware of this non-unique assignment. To visualize the overlap in the assignment of genes to chromatin states, I generated a Venn diagram of all genes that overlap with any of the repressive chromatin states 1, 3 or 5 (Figure 16B). Clearly, most genes that overlap with states 1 or 5 also overlap with state 3. Perhaps, this demonstrates that state 3 is to some extent a default state of repression, which is therefore found in inactive genes everywhere, even if those genes are additionally targeted by mechanisms that establish the other repressive states 1 or 5. In short, chromatin states 1, 3 and 5 are all repressive chromatin states, but their association with genes is not mutually exclusive.

To visualize the local chromatin state distribution I reproduced the profiles of the *eve* locus (Figure 7) but annotated the assignment of individual bins from the EM-model (Figure 17A). The chromatin state 5 (lime green) does indeed span the whole eve locus, as I predicted from the enrichment of H3K27me3. Around the locus, there are active genes, which have distinct chromatin states at the TSS (state 7, brown), enhancers (marked by H3K4me1, state 4, magenta) and the 3' gene body (state 6, yellow). To the right of the *eve* locus Pka-R2 is marked as belonging to chromatin state 3, which is well reflected by the presence of intermediate H3K27me3 levels as well as the absence of almost all active histone marks (except perhaps two small peaks of H3K4me1).

Figure 17: Chromatin states distribute across genes concordantly with the marks enriched in them. A: Genome profiles were produced as in Figure 7, but an additional track for the EM-clusters was added between polyA RNA and Gene annotation. B: Frequency of chromatin states relative to TSS. For all TSS the state assignment of the adjacent 10 bins upstream and downstream (= 2000 bp upstream and downstream) was determined and plotted as relative frequency over the number of total TSS.

This representation is however only a snapshot from a small genomic region. Therefore, I determined genome-wide the frequency of individual chromatin state assignments within the 21 adjacent bins centered around the TSS assignments (Figure 17B). The results show that also genome-wide the chromatin state assignments reflect the topology of active and repressed genes, with a pattern similar to the relative frequencies of ChIP targets around gene transcriptional start sites (Figure 10). Among the states that I determined to relate to active chromatin, state 7 has a major peak right after the TSS, with a minor peak slightly 5′ of it,

consistent with the high levels of H3K4me3 in that state (Figure 10). State 6, on the other hand, is enriched further downstream of the TSS, reflecting the tendency of H3K36me3 to be enriched on 3' ends of genes. State 6 is also enriched far upstream of the TSS, but this is likely to reflect signal deriving from adjacent genes.

Particularly interesting is the distribution of states 2 and 4, both of which I assumed to be enhancer chromatin states. These states seem to be most frequent directly 5' to the TSS, which is marginally different from the position of H3K4me1 I described earlier (Figure 10) and perhaps more closely resembles the location of a mixture of RNA polymerase II and H3K4me1. This proximity to the TSS possibly reflects a mixture of promoter and enhancer chromatin states close to the TSS others have reported.[277]

The repressive states of states 1, 3 and 5, on the other hand, show very little preference with regard to their distribution and are almost constant in their frequency around the TSS with only a slight decrease directly at the TSS. This is very similar to the distribution of repressive marks that I described earlier (Figure 10). In the mean frequency of all repressive states across all bins around the TSS repressive state 3 is much more frequent than the state 1 or 5, but this reflects the overall abundance of these states (Figure 13). Together, this shows that the distribution of chromatin states reflects the distribution of histone marks which they are enriched for.

In summary, using the binomial expectation-maximization algorithm, I implemented a novel approach to define chromatin states. After validating this algorithm based on published observations I conclude that the EM-based chromatin state model works comparably well to prior HMM-based models[55,69] while producing a mathematical model that is based on count based probability distributions found in ChIP-seq. Using this new approach, I was able to identify a previously not characterized repressive chromatin state, identified by the repressive of H3K27me3 at levels below those observed for developmentally regulated and stably repressed PcG chromatin. I reproduced this observation using a gene-based EM clustering approach, which further supports that two distinct H3K27me3 dominated chromatin states exist in *Drosophila* plasmatocytes. From here on, I will call the states derived from the 3 state gene-level H3K27me3 model H3K27me3 high, medium and low states and the 7 states from the combined ChIP-seq data chromatin states 1-7.

5.2.4 Plasmatocyte chromatin states and infection regulated genes

After determining both, a set of immune regulated genes and chromatin states in plasmatocytes, I tested which chromatin landscape is characteristic for inactive but still inducible genes. Therefore, I used the gene classifications based on the overlap with chromatin states (Figure 15)

and on the H3K27me3 gene state (Figure 12) and then applied them to the differential expression analysis from the plasmatocyte 3h, 6h and 18h post septic injury comparisons to control (Figure 3). First I tested enrichment of these states by applying the FRY algorithm implemented in limma, which is a fast approximation of the rotation based ROAST algorithm also implemented in limma.[278,279]

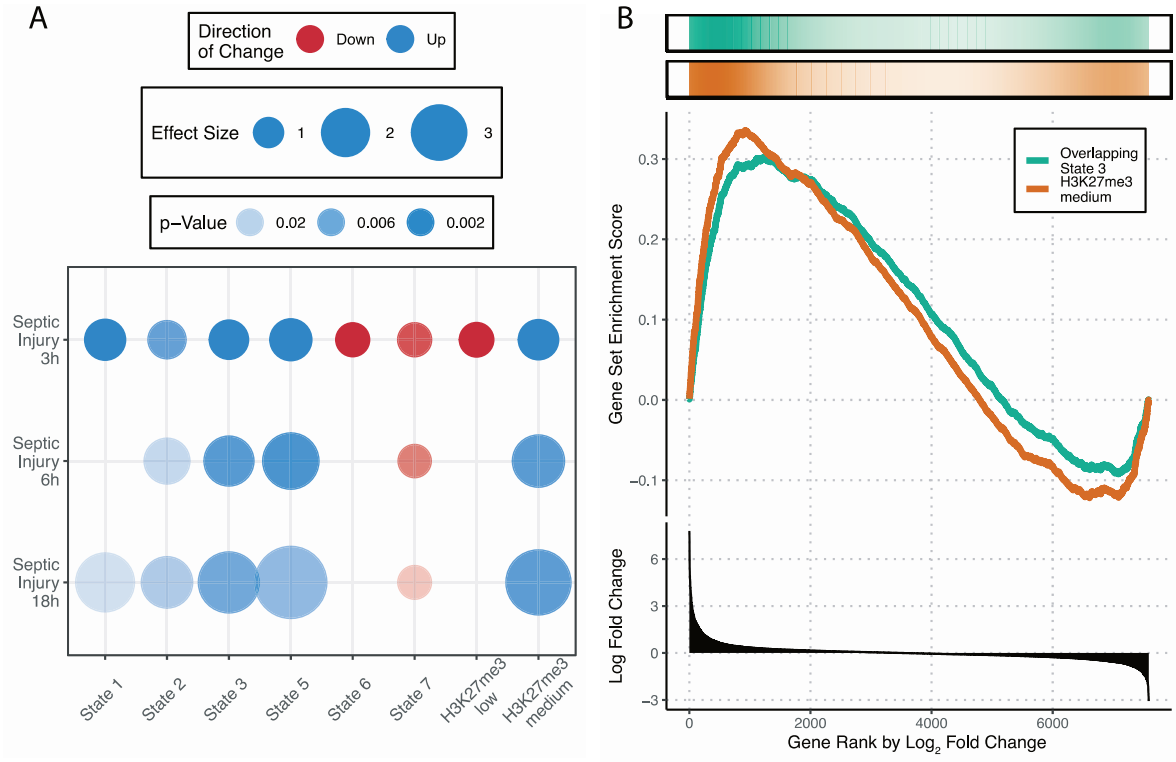

Figure 18: Plasmatocyte immune induced genes are enriched for repressive chromatin states. A: Differentially regulated genes in 3h, 6h, and 18h post septic injury plasmatocytes were tested for enrichment of overlapping chromatin states (Figure 15) and H3K27me3 level (Figure 12) using the limma fry algorithm. The color indicates if they were enriched among up- or down-regulated genes, size is the effect size (the relative fraction of genes overlapping with that state), and transparency the p-value. B: Gene set enrichment (GSEA) style plot of enrichment of genes overlapping chromatin state 3 (green) or falling into the H3K27me3 medium state (orange). Genes are ranked by fold change. Bottom panel: Log2 fold change over gene rank. Top panel: Density of genes positive for the feature. Middle: GSEA score by gene rank. P-value determined by gene permutation: Overlapping chromatin state 3: $p < 1*10^{-5}$, H3K27me3 medium state: $p < 1*10^{-5}$.

Figure 18A shows the enrichment that was observed for all the chromatin states and gene-based H3K27me3 levels, respectively, at the different time points after challenge. As with the GO-term analysis earlier, the size of the circle indicates the enrichment of the respective chromatin states

assignment, while the color indicates the directionality of the change and the transparency indicates the p-value of the chromatin states enrichment.

Both, chromatin state 4 and H3K27me3 high state groups are absent because they were not enriched in any of the comparisons. The data clearly indicates that chromatin states associated with active genes (chromatin states 6 and 7 as well as the H3K27me3 low state) are enriched among down-regulated genes, while up-regulated genes are enriched for repressive chromatin states (chromatin states 1,3 and 5 and H3K27me3 medium state genes). Further, all modules tend to have both smaller effect sizes and smaller p-values in the 3h post septic injury over control comparison than in the other comparisons. Likely the large number in overall differentially expressed genes in the 3h post septic injury comparison factors into this observation.

Interestingly, at 3h and to a much lesser extent at 18h post-challenge genes overlapping chromatin state 1 are enrichment. This enrichment is however smaller than the enrichment observed for the genes overlapping with the other 2 repressive chromatin states. Further, genes overlapping chromatin state 5 seems to have the strongest effect size at each of the time points, while genes overlapping chromatin state 3 have the smallest p-value. This may, in fact, be explained by some statistical properties relating to this enrichment analysis. First, the effect size is inversely correlated with the number of elements in the gene category. Since there are far fewer genes overlapping either chromatin state 1 or 5, their effect size is naturally larger. At the same time, a larger number of elements tend to produce smaller p-values, which explains why the genes overlapping with chromatin state 3 produce smaller p-values than those overlapping chromatin state 1 or 5. Beyond the statistical details of the enrichment analysis, it is important to recall that in my model genes can be assigned to multiple chromatin states simultaneously, in other words, there is no 'majority vote' that assigns a gene solely to the predominant chromatin state (Figure 16). Many genes overlapping with chromatin state 1 and 5 also overlap with chromatin state 3. Therefore the chromatin state assigned to some genes might actually not reflect the predominant chromatin state found at the respective gene. However, this does not apply to gene-level states of H3K27me3 (Figure 12), which is assigned uniquely for each gene. It is interesting, that in the gene enrichment analysis for up-regulated genes, of the 3-state H3K27me3 model only H3K27me3 medium state genes are overrepresented. The H3K27me3 high state is absent because no genes out of this group are differentially regulated. The comparison of gene-level enrichments for H3K27me3 with the bin signal levels showed H3K27me3 high state genes best correspond to chromatin state 5 (Figure 13). This suggests that the genes assigned to chromatin state 5 and up-regulated upon infection are likely those where this chromatin state is not the predominant one. Importantly, both the chromatin state-based as well as H3K27me3

gene level-based analysis show that genes located in H3K27me3 modified chromatin are enriched among the genes up-regulated upon immune challenge.

To ensure that these observations are not an artifact of the gene enrichment analysis method I tested if this enrichment is also found with an alternative analysis approach. Therefore I implemented a modified Gene Set Enrichment Analysis (GSEA)[280], a tool that is frequently used to test for gene module enrichment and visualize the results. Here genes are ranked by a ranking statistic, for which I chose the $\log_2$ fold change. Then a gene rank wise enrichment score is calculated by walking down the ranked gene list and increasing the score, if the gene that is met belonged to the category of interest, and else decreasing it. The maximum enrichment score can then be used for permutation-based statistical testing. Figure 18B shows the GSEA for genes overlapping with chromatin state 3 (green) and H3K27me3 medium state genes (orange). The bottom panel shows the ranking statistic (the $\log_2$ fold change), while the top panel shows the density of genes that are assigned to the category of interest at that gene rank (overlapping with chromatin state 3 or H3K27me3 medium state). The middle panel shows the enrichment score derived from the analysis of this density as discussed above. Both terms, overlapping with chromatin state 3 and H3K27me3 medium state, show strong enrichment among genes with a high log-fold change, and both reach enrichment scores of greater than 0.3. To test the significance of these findings, I applied gene-wise permutation to them. For neither gene set a higher gene set enrichment score was found after 100.000 permutations, resulting in $p < 1*10^{-5}$ for both sets.

This underscores that the H3K27me3 intermediate chromatin, with H3K27me3 levels lower than that of developmental genes, but clearly above background, marks inactive but inducible genes. This observation is in contrast to the hypothesis that H3K27me3 leads to stable repression of targeted loci.[25,281]

To quantify H3K27me3 enrichment levels at immune regulated genes I categorized the RNA-seq data set into up-regulated, not significantly different, down-regulated, or not expressed genes and compared these to H3K27me3 enrichment (as calculated by (18)). As described earlier, I defined genes not expressed if their read count appeared insufficient for statistical inference, that is to say, if they had less than 1 read mapped to them per 1 million reads in a sample. Differential expression was defined by both a false discovery rate corrected p-value < 0.05 and a $\log_2$ fold change > 1 (or < -1). I calculated gene-level enrichment of H3K27me3 according to (18), formed the mean between replicates and plotted this enrichment for all genes (Figure 19). Figure 19A recapitulates that for 6h post septic injury over control H3K27me3 intermediate genes are overrepresentation. Both down-regulated genes and genes which are not differentially

expressed contain almost exclusively H3K27me3 low genes, which based on the earlier gene-level EM chromatin state (Figure 12) have an enrichment of less than 0.4. There are however slightly higher levels of H3K27me3 in the down-regulated gene group ($p = 0.0045$, Wilcoxon rank-sum test). The box-plots demonstrate that up-regulated genes are much higher than not differentially expressed genes in H3K27me3 enrichment ($p < 2.2*10^{-16}$, Wilcoxon rank-sum test). The underlying distribution (violin plots) suggests, however, that this is due to a mixture of H3K27me3 negative genes and H3K27me3 positive genes. Among up-regulated genes there is also a number of H3K27me3 low genes (enrichment < 0.4). There is however a second population, a large fraction of which falls into the category of genes with enrichments between 0.4 and 0.75, which are the H3K27me3 medium state genes (compare Figure 12). Therefore, up-regulated genes still are much lower in H3K27me3 enrichment than not expressed ($p < 2.2*10^{-16}$, Wilcoxon rank-sum test), where a large fraction of genes are in the H3K27me3 medium state or H3K27me3 high state (all with enrichment > 0.4). In summary, a set of H3K27me3 positive genes is up-regulated in plasmatocytes after septic injury, with a stronger enrichment for such genes at later time points post septic injury.

I was also interested in how this enrichment behaves quantitively over time, in other words what the relationship between unchallenged plasmatocyte H3K27me3 levels and the dynamics of gene expression after immune challenge is. I showed earlier (Figure 18A) that at 3h post septic injury, the p-value of H3K27me3 positive gene enrichment was smaller than at either 6h or 18h after septic injury, while the effect size was larger at both 6h and 18h than at 3h after septic injury. I, therefore, selected from each comparison (3h, 6h and 18h post septic injury vs control) the up-regulated genes and calculated their H3K27me3 enrichment. Interestingly, H3K27me3 enrichment is higher at later time points (3h vs 6h $p = 3.4*10^{-14}$, 6h vs 18h $p = 0.0013$, Wilcoxon rank-sum test) (Figure 19B). This means that genes that are up-regulated early after infection rarely come from H3K27me3 marked chromatin but genes that are up-regulated later or stay on longer are more likely to come from an H3K27me3 positive chromatin state. There is a clear bimodal distribution in the H3K27me3 enrichment levels, with both H3K27me3 low (enrichment < 0.4) and H3K27me3 medium state (enrichment > 0.4 and < 0.75) genes being present in all conditions. In fact, the change in H3K27me3 enrichment levels seems to derive from a shift in ratios between H3K27me3 low and H3K27me3 medium state genes. It is, therefore, important to consider the vastly different total number of up-regulated genes (3h: 1311 genes, 6h: 187 genes, 18h 56 genes). Hence, it is not clear from this observation alone, whether H3K27me3 positive genes are up-regulated later, H3K27me3 negative genes are down-regulated earlier, or if another

effect is the cause of the observation that the fraction of H3K27me3 intermediate genes increases over time.

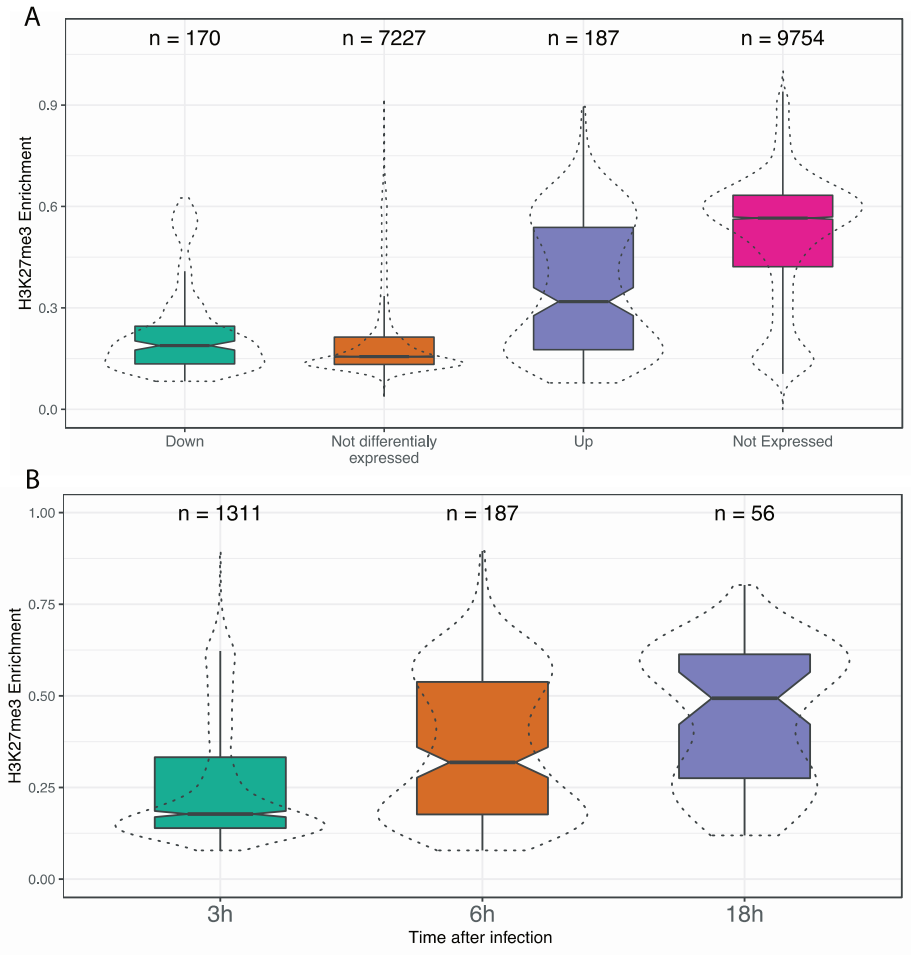

Figure 19: H3K27me3 levels are increased in immune induced genes. H3K27me3 enrichment was calculated for all genes using the mean enrichment from both replicates. Genes were grouped into categories based on their differential expression in plasmatocytes after challenge. A: Genes were grouped by their expression at 6h post septic injury when compared to control, either down-regulated (Down), not significantly differentially expressed or up-regulated (Up). Not expressed marks genes that were filtered by their low expression and have less than 1 read per million mapped to them. n: number of genes in that group. Mann-Whitney *U* test: Up vs not differentially expressed $p < 2.2*10^{-16}$, Up vs Not expressed $p < 2.2*10^{-16}$, Down vs not differentially expressed $p = 0.0045$. B: For each time point of septic injury, up-regulated genes at that time when compared to control were selected and plotted as described for A. Mann-Whitney *U* test: 3h vs 6h $p = 3.4*10^{-14}$, 6h vs 18h $p = 0.0013$.

Consequently, I took all genes which were up-regulated at 6h post septic injury and split them based on their gene-level assignment to H3K27me3 (high + medium vs. low) or based on their chromatin state assignment (state 3 vs. all other states). From these genes, I plotted the $\log_2$ fold changes over time when compared to the unchallenged control (Figure 20). Interestingly, for

neither grouping based on H3K27me3 nor on overlap with chromatin state 3, there appears to be much of a difference in the overall shape of the response between H3K27me3 positive and negative genes. Both groups are increased in expression immediately at 3h, stay at similar levels through 6h and then decrease slightly towards 18h.

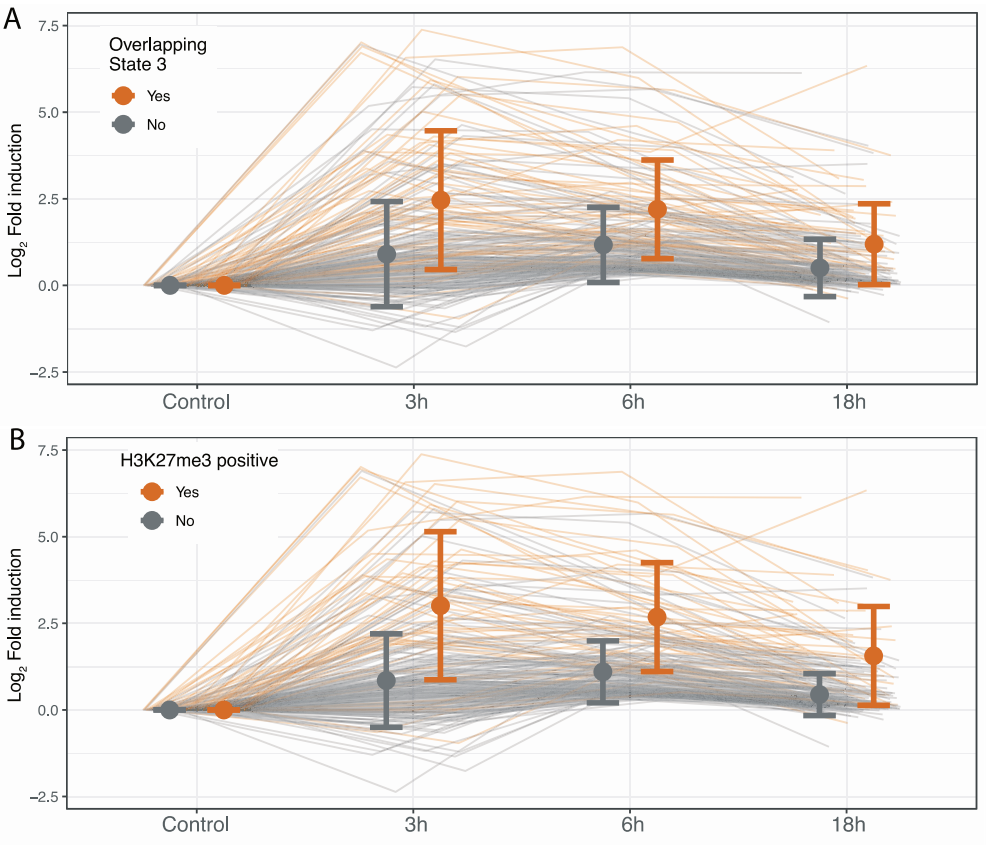

Figure 20: Genes in repressive chromatin states are not delayed in their transcriptional response. Log_2 fold changes for all genes which are up-regulated at 6h post septic injury is plotted as time course, grouped by their A: H3K27me3 level or B: their overlap with the H3K27me3 intermediate chromatin state 3. Orange genes fall in the respective category, while gray genes do not. Background traces are time courses of individual genes, foreground points show mean and standard deviation. (H3K27me3 high vs low, 3h: $p = 6.5*10^{-12}$, 6h: $p = 2.6*10^{-15}$, 3h: $p = 5.8*10^{-11}$; Overlapping chromatin state 3 vs Not, 3h: $p = 2.9*10^{-10}$, 6h: $p = 4.6*10^{-11}$, 3h: $p = 2.4*10^{-8}$, Wilcoxon rank-sum test)

The magnitude of the fold change on the other hand is much greater for the H3K27me3 positive genes at all time points and for all comparisons (H3K27me3 high vs low, Figure 20B, 3h: $p = 6.5*10^{-12}$, 6h: $p = 2.6*10^{-15}$, 3h: $p = 5.8*10^{-11}$; Overlapping Chromatin state 3 vs Not, Figure 20A, 3h: $p = 2.9*10^{-10}$, 6h: $p = 4.6*10^{-11}$, 3h: $p = 2.4*10^{-8}$, Wilcoxon rank-sum test). Hence, H3K27me3 positive genes do not respond delayed when compared to H3K27me3 negative genes. Neither do they return to a base level significantly faster. Together, this indicates that H3K27me3

positive genes are not slower in responding when compared to H3K27me3 low genes, but their response may be stronger than that of H3K27me3 low genes.

In summary, I have shown that H3K27me3 positive genes are up-regulated in plasmatocytes after immune challenge. These inducible genes fall into an H3K27me3 intermediate category, that is distinct from the H3K27me3 high category which includes for example stably repressed developmental genes and a H3K27me3 depleted category which includes actively transcribed genes. These H3K27me3 positive immune regulated genes are a subset of all immune regulated genes, they are robustly induced after immune activation and are not delayed in their transcriptional response. Therefore, I hypothesized that H3K27me3 is targeted to immune genes to prevent their aberrant transcription and thereby acts as an immune regulator in *Drosophila*.

5.3 Differential ChIP-seq of challenged plasmatocytes

5.3.1 ChIP-seq of 6h post septic injury plasmatocytes

The mere presence of H3K27me3 at immune inducible gene loci does not prove that H3K27me3 acts as a suppressor of transcription at these genes. However, if it did act to suppress transcription at these genes then it should be removed to allow their transcriptional activation.

Therefore, I used ChIP-seq to test for differences in histone modifications after immune challenge. Because ChIP-seq required a larger number of cells than RNA-seq, here I only tested one time point at the 6h after septic injury. This 6h time point has been widely used in the literature as a time point to identify transcriptional changes after challenge.[205,225,232] Further, the RNA-seq of plasmatocytes revealed a significant number of H3K27me3 positive genes differentially expressed at this time point. Therefore, I collected cross-linked chromatin samples from plasmatocytes 6h post septic injury and performed ChIP-seq on them with antibodies directed against H3K27me3 and H3K27ac. It seemed plausible that H3K27me3 might be accompanied by a gain of H3K27ac, because H3K27ac is a mark found at actively transcribed genes and because this modification is mutually exclusive and therefore competing with H3K27me3.[42] As described for the earlier ChIP-seq the samples were checked for general enrichment by qPCR and ChIP-seq libraries were prepared from them. These libraries were then sequenced and mapped to the BDGP dm6 reference genome exactly the same way as all ChIP samples from unchallenged plasmatocytes.

I expect that, while there will be treatment depended changes in histone modifications, they are likely to be minor compared to the overall ChIP-seq signal given that the plasmatocytes are neither transdifferentiated nor have they changed in any other fundamental way. Additionally, all histone modifications have a characteristic distribution, that does not change fundamentally even when comparing ChIP-seq from entirely different cells. I therefore expected, that the samples of the same type (H3K27me3, H3K27ac or input) are very similar not only between replicates but also between unchallenged and 6h post septic injury samples.

Hence, I first applied PCA to check for sample similarity. Accordingly, I split the genome into 10 kb bins and quantified the signal by counting reads that fall into them. I then picked the 1000 most variant bins and applied PCA, which is plotted in Figure 21.

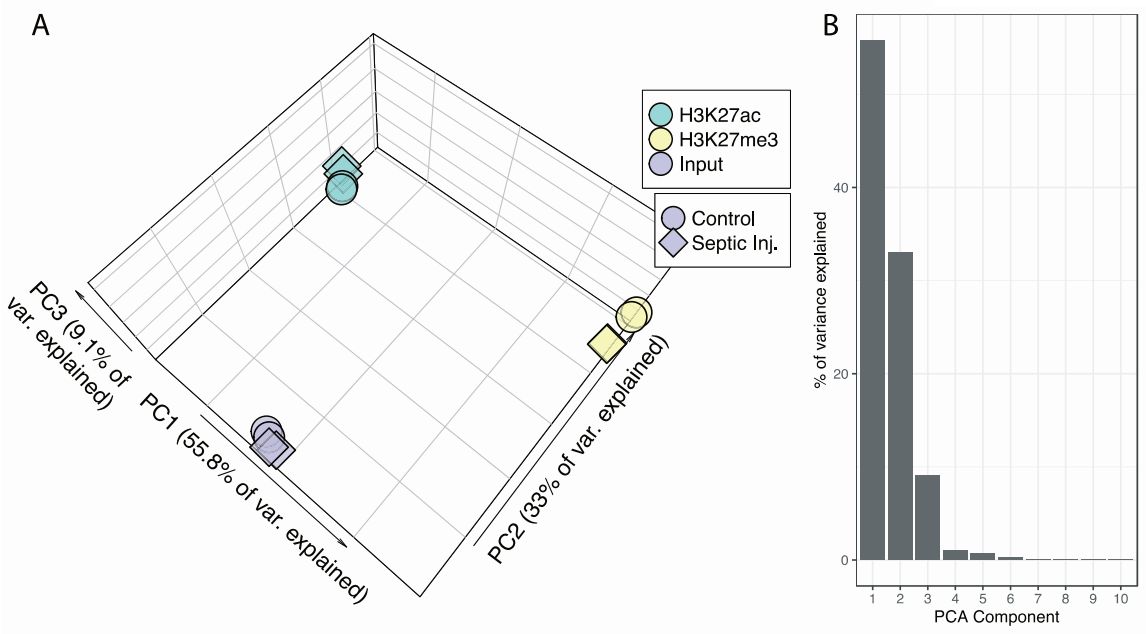

Figure 21: ChIP-seq from plasmatocytes cluster primarily by the histone modification and not the treatment. The genome was split into bins of 10 kb and reads falling into each bin were counted. The resulting counts were subjected to principal component analysis. A: Visualization of the first 3 components of the PCA. Colors indicate antibody targets, shapes indicate plasmatocyte treatments. B: Percent of total sample variance explained by the first 10 components.

When plotting the first 3 components of the PCA (Figure 21A) indeed all samples cluster according to the antibody target (H3K27me3 in yellow, H3K27ac in green, input in purple) rather than according to treatment condition (unchallenged control as a circle and 6h post septic injury as a square). This suggests that the changes in signal introduced by the septic injury only make up a minor part in the overall signal and that the plasmatocyte specific chromatin landscape is still the main cause of the observed ChIP enrichment. Further, similar to unchallenged plasmatocytes, the active mark H3K27ac, and the repressive mark H3K27me3 separate strongly along the principal component 1, which explains 55.8% of the total sample variance, while the input sample lies in between the two. Interestingly, component 2, which explains 33% of sample variance separates both ChIP samples from the input, which likely represent regions where neither of the modifications are enriched. When I compare for each of the ChIP targets the sequencing results derived from control plasmatocytes to 6h post septic injury plasmatocytes they cluster slightly apart, an observation which is perhaps most pronounced in the H3K27me3 ChIP, where the distance between the two control samples and the two septic injury samples is largest. There is also an interesting directionality of change: In H3K27me3 the septic injury samples are slightly closer to the input samples than the control samples, while the H3K27ac samples show the reverse effect. This may be a first indication, that H3K27me3 is lost in some

regions, corresponding to a greater similarity to the modifications free like input, while H3K27ac is increased in distance to input, indicating a gain of the signal. Therefore, ChIP-seq samples of unchallenged and 6h post septic injury plasmatocytes cluster closely together by the immune-precipitation target and only separate weakly by the treatment.

Hence, I analyzed the changes in the ChIP signal as well as the location of those changes with regard to the genes of interest, which are immune induced genes. One of the established procedures to check for differences between ChIP-seq samples employs MACS2[258,260]. Here the identification of differentially targeted regions is done in several steps. First, all regions with higher than background enrichment are identified across all conditions of the ChIP-seq samples in a step named peak calling. Given the broader binding of histone modifications I here used the MACS2 broad peak algorithm. Then those peak regions which differ in signal intensity are identified by determining the probability of the read counts in peak regions deriving from the same Poisson model across the different treatment conditions. I applied this pipeline for both H3K27me3 and H3K27ac separately. Then I used these peak calls to identify differentially targeted genes, which I defined as any gene that overlaps with at least 1 bp with a differential peak region of that ChIP-seq target.

In addition, I applied a differential binomial EM approach based on the algorithm proposed in the normR package[264]. Here, instead of calling the probabilities of ChIP over input as described in Figure 9, I identify different populations as if the reads were drawn from urns containing mixed reads from the control and septic injury ChIP-seq runs. From these enriched regions, I can identify genes with differential histone modifications by identifying overlap of differentially modified regions and genes.

At first, I tested this pipeline on the set of H3K27me3 ChIP-seq data sets (Figure 22). Using MACS2 a large number of broad peak regions can be identified in the control plasmatocyte samples (12756 peaks, Figure 22A), of which only a small fraction (1210 peaks) contains regions that are differentially modified. This is in line with the fact many H3K27me3 positive genes are not related to immune function and should be unaffected by the septic injury challenge. It is interesting that all of the H3K27me3 peak regions which are changed between control and septic injury conditions have lower levels of H3K27me3 after septic injury, while none show increased levels. This is exactly what I would expect, if H3K27me3 is a regulator of immune gene transcription since the loss of H3K27me3 would enable increased transcription of the overlapping genes. At the same time, there is no reason to believe that down-regulation of genes would immediately result in H3K27me3 modification of those genes, which explains why no regions were found to be increased in H3K27me3 after septic injury.

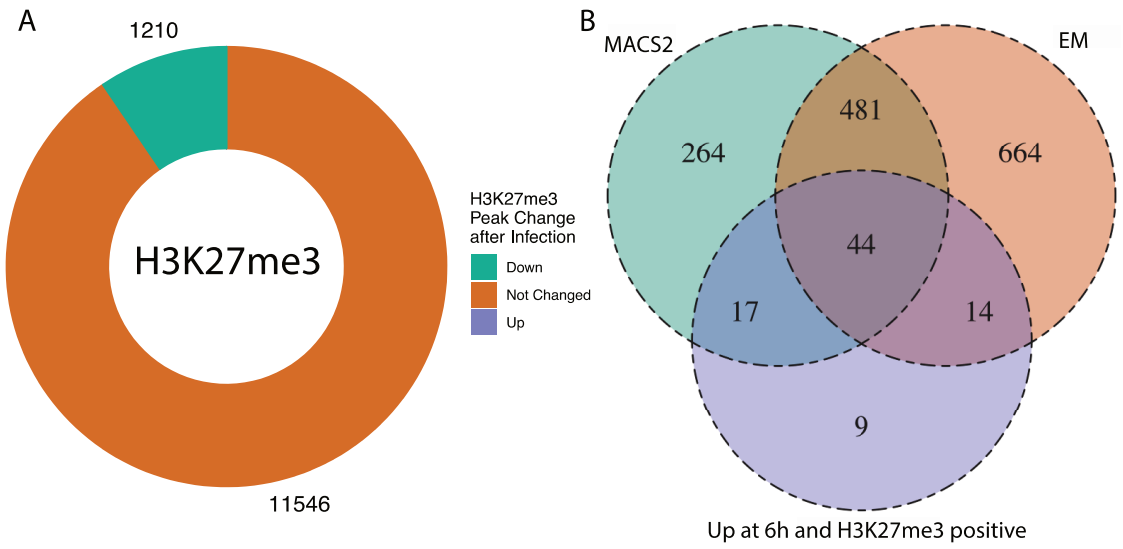

Figure 22: H3K27me3 peaks are reduced in H3K27me3 in plasmatocytes at 6h post septic injury. Differentially bound regions for H3K27me3 were calculated using MACS2 and a differential EM approach. A: Donut plot of all peaks called in non-challenged plasmatocyte H3K27me3 sample using MACS2. Colors indicate change between6h post septic injury and unchallenged plasmatocytes. B: Venn diagram of genes overlapping MACS2 H3K27me3 differential peaks (MACS2), EM model H3K27me3 differential peaks (EM) and genes differential expressed at 6h and high in H3K27me3

I then compared how these changes relate to the genes I found differentially expressed after septic injury. I therefore plotted three sets of genes in a Venn diagram: (1) Genes identified by MACS2 to be in regions that change H3K27me3 enrichment, (2) genes identified by the binomial EM algorithm to be in regions that change H3K27me3 enrichment and (3) genes which were both differentially expressed at 6h after septic injury as well as H3K27me3 positive (Figure 22B). First, the overlap between MACS2 and EM with regard to identifying genes that gain or lose H3K27me3 is good. The binomial EM algorithm is calling more genes increased in H3K27me3 but covers more than two-thirds of all genes called by MACS2. More importantly, however, between both algorithms, they cover 75 out of the 84 genes that are H3K27me3 positive and up-regulated at 6h post septic injury. This high level of overlap occurs in spite of the fact, that I applied stringent false discovery rate corrected p-value thresholds, which means there is likely a number of false negatives for both RNA-seq and ChIP-seq changes. Therefore, the observed overlap between differentially expressed genes and gene regions that loose H3K27me3 supports well the hypothesis that H3K27me3 is lost at these genes in plasmatocytes after immune activation.

Additionally, I applied the same analysis to the H3K27ac samples. Using the same MACS2 peak calling approach as described for H3K27me3, I find that H3K27ac is increased in many of the peak regions at 6h post septic injury when compared to control plasmatocytes (Figure 23A). Here, 545 peak regions gain H3K27ac enrichment and only 10 peak regions are decreased in enrichment. This is in contrast the H3K27me3 ChIP-seq results and therefore well complements the data on H3K27me3 loss and gene up-regulation.

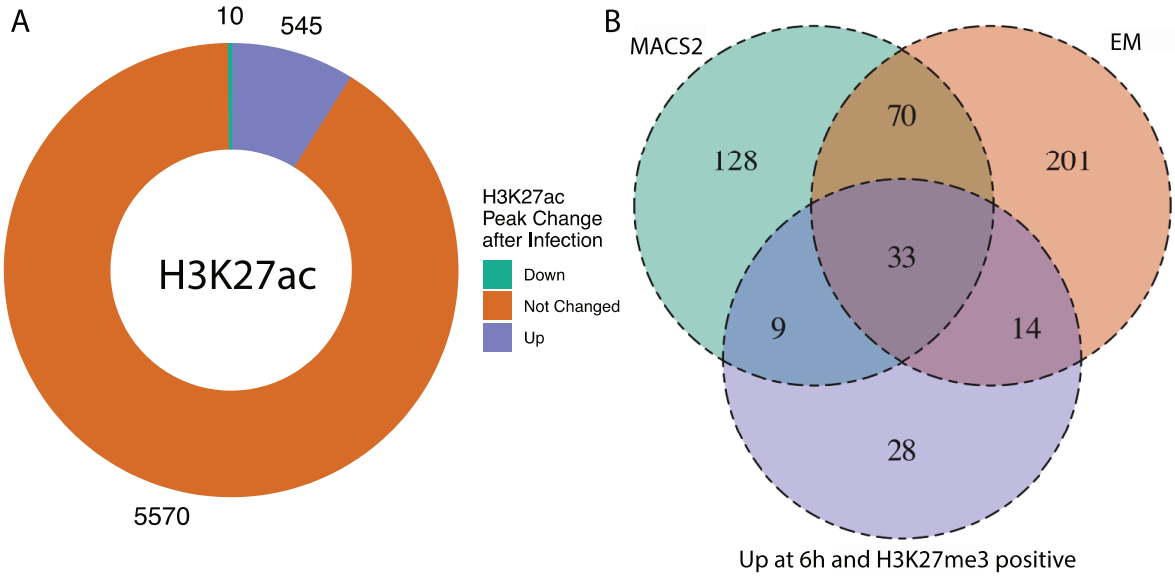

Figure 23: H3K27ac peaks are increased in H3K27ac in plasmatocytes at 6h post septic injury. Differentially bound regions for H3K27ac were calculated using MACS2 and a differential EM approach. A: Donut plot of all peaks called in non-challenged plasmatocyte H3K27ac sample using MACS2. Colors indicate change between 6h post septic injury and unchallenged plasmatocytes. B: Venn diagram of genes overlapping MACS2 H3K27ac differential peaks (MACS2), EM model H3K27ac differential peaks (EM) and genes differential expressed at 6h and high in H3K27me3

Subsequently, I identified genes associated with differentially called H3K27ac peaks by the MACS2 and the binomial EM algorithms and plotted their overlap with genes that were H3K27me3 positive and up-regulated at 6h after septic injury in a Venn diagram (Figure 23B). Compared to H3K27me3, for H3K27ac the different groups overlap less well. The MACS2 and EM algorithms only show a less than half overlap. Further, they together only cover 56 out of the 84 H3K27me3 positive immune induced genes. However, an increase in H3K27ac may very well also be unrelated to the loss of H3K27me3 I observed. While it is true that H3K27ac can act as a competitor for trimethylation at histone H3 lysine 27, it may also be placed by completely unrelated mechanisms, perhaps on genes that are differentially regulated, but not necessarily repressed by H3K27me3 in unchallenged plasmatocytes. Additionally, there is no reason to believe that H3K27ac is actively required to induce transcription of H3K27me3 positive genes, and

it may either be established with some delay or not at all. Last, H3K27ac is found both, 3' to the TSS and at enhancers and regulatory elements, but whether the role of H3K27ac is the same at every location is not yet clear. Together, this indicates that H3K27ac is induced at a number of genes in plasmatocytes after septic injury, but this increase is not restricted to genes in H3K27me3 positive gene states.

Therefore, I checked if this overlap of differentially modified regions and differentially expressed genes is also represented in the enrichment of differentially modified genes among up-regulated genes at 6h post septic injury. Here, to visualize the data, I used the same gene set enrichment analysis algorithm I described before (Figure 18): Genes are sorted based on a ranking statistic (here the $\log_2$ fold change) and then a rank level enrichment score is calculated by walking down the list and adding or subtracting from the score, depending on what type of gene is met. This analysis I applied to the 4 differential ChIP-seq gene sets: Genes called differential by either the MACS2 or the EM algorithm for either H3K27me3 or H3K27ac (Figure 24).

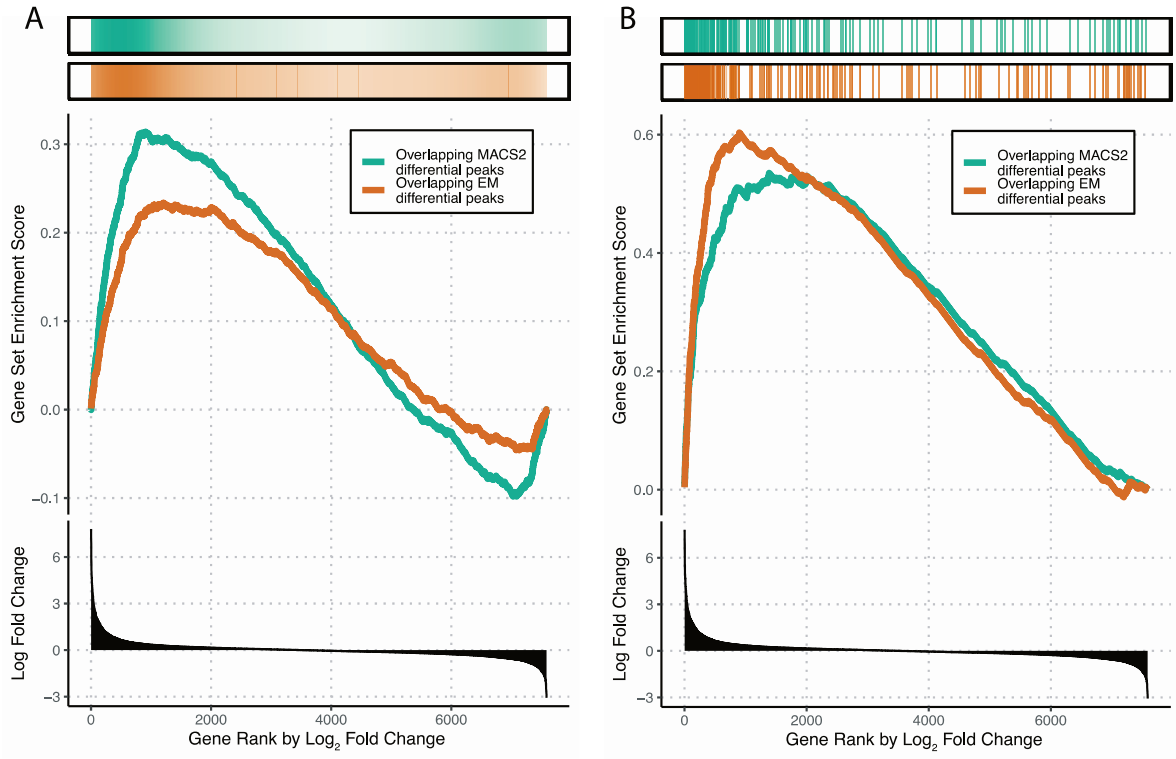

Figure 24: Genes with differential H3K27me3 and H3K27ac are enriched among immune induced genes in plasmatocytes. Genes overlapping differential peaks in H3K27me3 (A) or H3K27ac (B) were determined by either MACS2 or differential EM. Gene set enrichment was then performed for those gene groups using the 6h post septic injury RNA-seq. Bottom panels: Log2 fold change of 6h post septic injury vs. control over gene rank. Top panels: Density of genes positive for the feature. Middle: GSEA score by gene rank. P-values

by fry algorithm: H3K27me3 differential by MACS2 p = 0.0088, EM p = 0.0118. H3K27ac differential by MACS2 p = 0.0005, EM p = 0.0006. P-values by GSEA permutation: all $p < 1*10^{-5}$

For the H3K27me3 data set (Figure 24A), the top panel shows the density of genes which were called as differential by either algorithm based on the gene rank, while the middle panel shows the rank-based gene set enrichment score. Both MACS2 and EM show a good enrichment among the up-regulated genes, such that permutation-based p-values are low ($p < 1*10^{-5}$ for both comparisons). Further, the enrichment for the gene set that includes genes identified by MACS2 is enriched slightly more, perhaps representing the overall smaller number of genes in this set when compared to the set identified by the EM algorithm (also see Figure 22B). Concordantly, applying the fry algorithm I find p-values of p = 0.0088 for MACS2 and p = 0.0118 for EM differential peak enrichment. For the H3K27ac differential genes, there is a similarly significant enrichment (Figure 24B). Here there is a strong enrichment among up-regulated genes with enrichment scores even larger than the ones observed for H3K27me3. Permutation-based p-values are small ($p < 1*10^{-5}$ for both comparisons), as were p values calculated by the fry algorithm (enrichment for genes that overlap with regions called differential by MACS2: p = 0.0005, or EM: p = 0.0006). Therefore, genes with differential H3K27me3 peaks and with differential H3K27ac peaks are enriched among genes that are up-regulated at 6h post septic injury in plasmatocytes.
It should also be possible to directly compare signals of ChIP-seq quantitively, much in the same way that it is done in ChIP-qPCR experiments. Therefore, I characterized the samples better to identify what data transformations are necessary. First, I determined Pearson correlation coefficients of samples in the same way I did for the ChIPs from unchallenged plasmatocyte. I divided the genome into 10 kb bins, counted overlapping reads and calculated Pearson correlation coefficients. These were sorted by hierarchical clustering and plotted as a heatmap (Figure 25A).
Similar to before (see Figure 8), the data sets correlate well between replicates, with correlation coefficients ranging between r = 0.98 and r = 1. ChIP-seq of different targets do not produce good correlation, with H3K27me3 and H3K27ac being very dissimilar, as I described for the ChIP-seq from unchallenged plasmatocytes. Further information can be gained by comparing the correlation coefficients of 6h post septic injury (infected) and control (resting) samples for the same antibody. For all samples (even the input) replicates of the same condition correlate slightly better than the different treatment conditions. This is likely indicative of those regions which are differentially modified between unchallenged and 6h post septic injury plasmatocytes. Particularly for H3K27me3 the two samples from 6h post septic injury plasmatocytes correlate slightly better with each other than they do with the control samples. Given that my previous

results revealed clear signal differences between 6h post septic injury and control samples, I wondered if technical factors of the ChIP-seq affect the samples and partially hide the true signal.

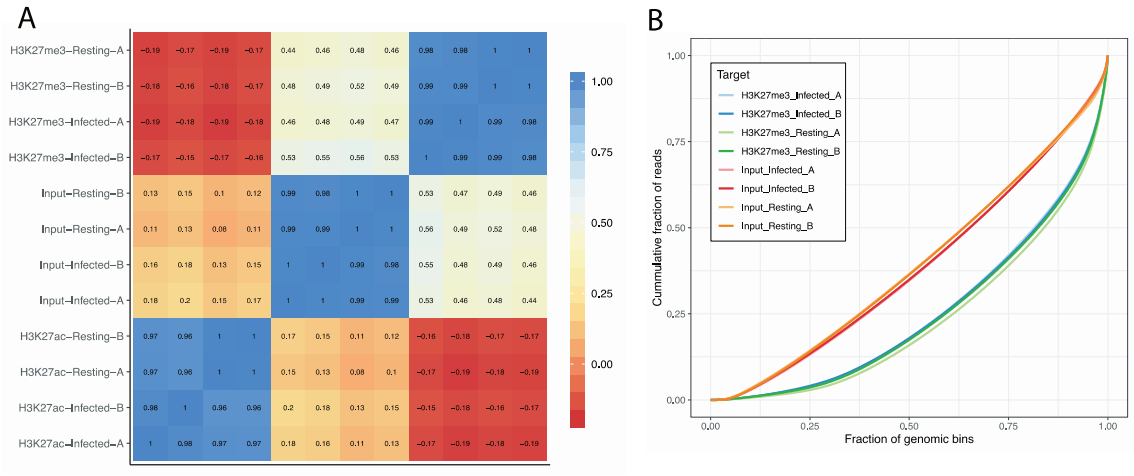

Figure 25: ChIP-seq replicates comparison shows low-level technical variance across samples. A: Mapped reads from each individual ChIP-seq replicate were binned using 10kb windows. Pairwise Pearson correlation coefficients of the resulting read counts were calculated using Deeptools. Samples were sorted using hierarchical clustering and visualized as a heatmap. B: ChIP fingerprint plot. Evenness of coverage was determined by calculating the cumulative fraction of reads that are contained in each quantile of genomic bins.

Therefore I generated fingerprint profiles[282] from the H3K27me3 ChIP-seq samples. To do this, the genome is divided into small bins (in this case 25 bp) and the read count is determined for each bin. Then the bins are sorted based on their read count in ascending order and the cumulative sum of the values is calculated. The resulting profiles can then be used to infer about the enrichment of the ChIP-seq: In ideal input samples, this profile would essentially be a diagonal line, since every bin is covered by the same number of reads, and therefore the cumulative fraction of reads at every bin is just the cumulative fraction of bins. This ideal input is impossible, because of random fluctuations in read counts and DNA copy number variations between *Drosophila* lab strains and the reference genome assemblies of BDGP dm6. For ChIP samples on the other hand, in the most extreme case, the profile would be zero for most of the fraction of bins, since they are not precipitated due to the lack of a given histone modification, and the profile would then quickly rise to 1 in the last few bins which are regions that actually carry that histone PTM. At the same time, ChIP samples always have background from non-specific precipitation, and some modifications are broadly spread across the genome. Therefore, any true ChIP profile will be somewhere in between this ideal input and the ideal ChIP. Figure 25B shows

these fingerprint profiles for H3K27me3 and input samples. The input samples are much closer to the diagonal than the H3K27me3, which is expected to have many enriched regions across the genome. But while all inputs and all H3K27me3 ChIPs run very close together, at least one of the H3K27me3 samples is slightly offset. This cannot be explained by biology differences, since most genomic regions should be unaffected by the septic injury, and even more, replicates should definitely not differ in their enrichment. It seems therefore likely, that there is a difference between immune-precipitations of the same histone target which derive from effects of slightly more rigorous washing steps, or slight changes in the ratios of antibody to chromatin, or other technical variations. Hence. these differences are not biologically meaningful and should, therefore, be normalized before further data analysis.

Indeed, when comparing the enrichment using 25 bp bins of the two replicates of unchallenged plasmatocytes H3K27me3 ChIP (Figure 26A), there is a bias in the distribution such that replicate B produces higher values for the same bins. The bin-wise enrichment values from this analysis are distributed in a banana-like shape, indicating that in replicate B the immunoprecipitation was more efficient in precipitating genomic regions with intermediate enrichment, relatively to the precipitation of both H3K27me3 low and high regions.

Since this distortion was not biologically meaningful but interferes with quantification, I decided to normalize the data using a quantile normalization approach. These quantile normalization methods have been well studied and used extensively in microarray-based genome analysis, where they are applied to filter out technical variations caused for example by differences in labeling efficiencies.[283,284] These technical variations in microarray signals are very similar to the ones I observed for my ChIP-seq replicates, so I applied a quantile normalization algorithm. This algorithm follows 3 steps: First, the genomic bins from each replicate are separately sorted based entirely on their read count value, irrespective of that bins genomic position. Then to each bin, the mean of all bins at the same sorted rank across replicates is assigned. Last the list is sorted back according to the genomic position of the bin. The exact code can be seen at Code Snippet 3. There are however some assumptions that need to be justified in order for quantile normalization to be appropriate. First, it assumes that the total abundance of H3K27me3 is the same. Since quantile normalization assigns each bin the row-wise mean, that also means that the sum of all bins, which is equivalent to the total signal across all the genome, will appear to be the same between all samples. Here I think this approximation is warranted. Because only a small subset of genes is changed in response to the septic injury, this would likely only cause a minor perturbance in the total levels of H3K27me3 in the plasmatocytes. This is also demonstrated by the fact that the majority of H3K27me3 peak regions are unchanged (Figure 22A). Second, in

sparse distributions, outliers will be affected to a much greater extend by changes. This is a large concern in microarrays, where few genes like for example *Actin* genes represent a significant fraction of the total mRNA. Here I used the H3K27me3 enrichment, which I calculated according to (18). Since in this definition all observed values must be between 0 and 1, which puts both an upper and lower limit to enrichment and prevents extreme values, the risk that quantile normalization would here cause distortions in ChIP-seq enrichment is minimal.

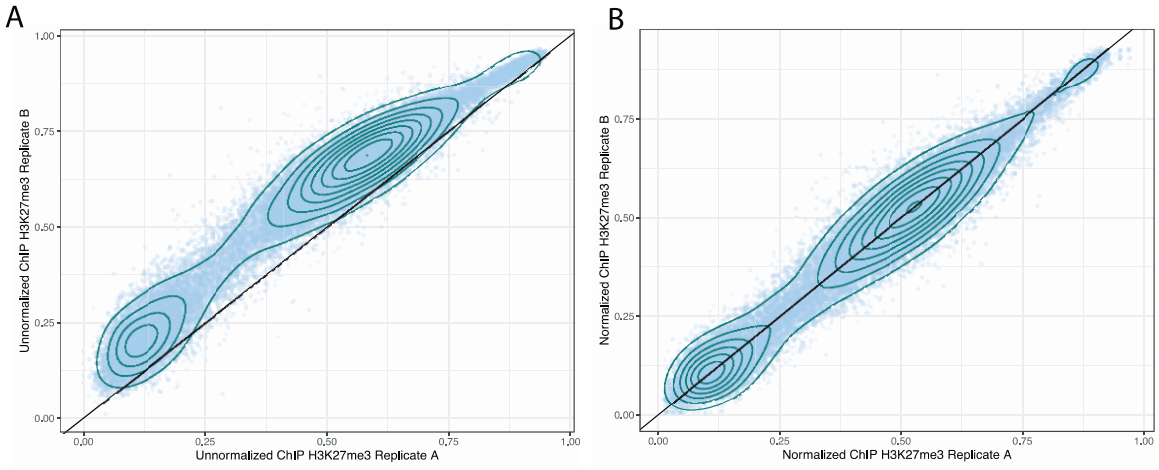

Figure 26: Quantile normalization eliminates technical variance in ChIP-seq replicates. A: H3K27me3 enrichment was calculated for 25 bp genomic bins of both replicates of H3K27me3 in unchallenged plasmatocytes. The outlines show point density, the diagonal line shows a 1:1 correlation (ideal model). B: The Enrichment of A was processed by a quantile normalization algorithm.

When applying this quantile normalization to the previously 'banana' shaped distribution of H3K27me3 enrichment of unchallenged plasmatocyte replicates, the signal is afterwards distributed along the diagonal, as I would expect from replicates of the same plasmatocyte condition (Figure 26B). At the same time, many important features are conserved: The data is still distributed in a trimodal fashion which has two major peaks at small and medium values and a minor peak at large enrichment values. Also, the diameter of the distribution orthogonal to the diagonal is unchanged, indicating that I only adjusted for technical biases, but did not suppress biological signal or variance. Hence, I applied this quantile normalization to both the H3K27me3 and H3K27ac enrichment data between the 4 samples targeting the same antigen and used this data for all following comparisons of differential ChIP-seq.

First, I applied the normalized data to visualize ChIP-seq signals. Accordingly, I quantile normalized the 25 bp window ChIP tracks and loaded them into IGV. In addition, I calculated the difference between the control and 6h post septic injury tracks of both H3K27me3 and H3K27ac,

to more easily identify regions with an increase or decrease in signal. Two examples from these tracks are shown in Figure 27.

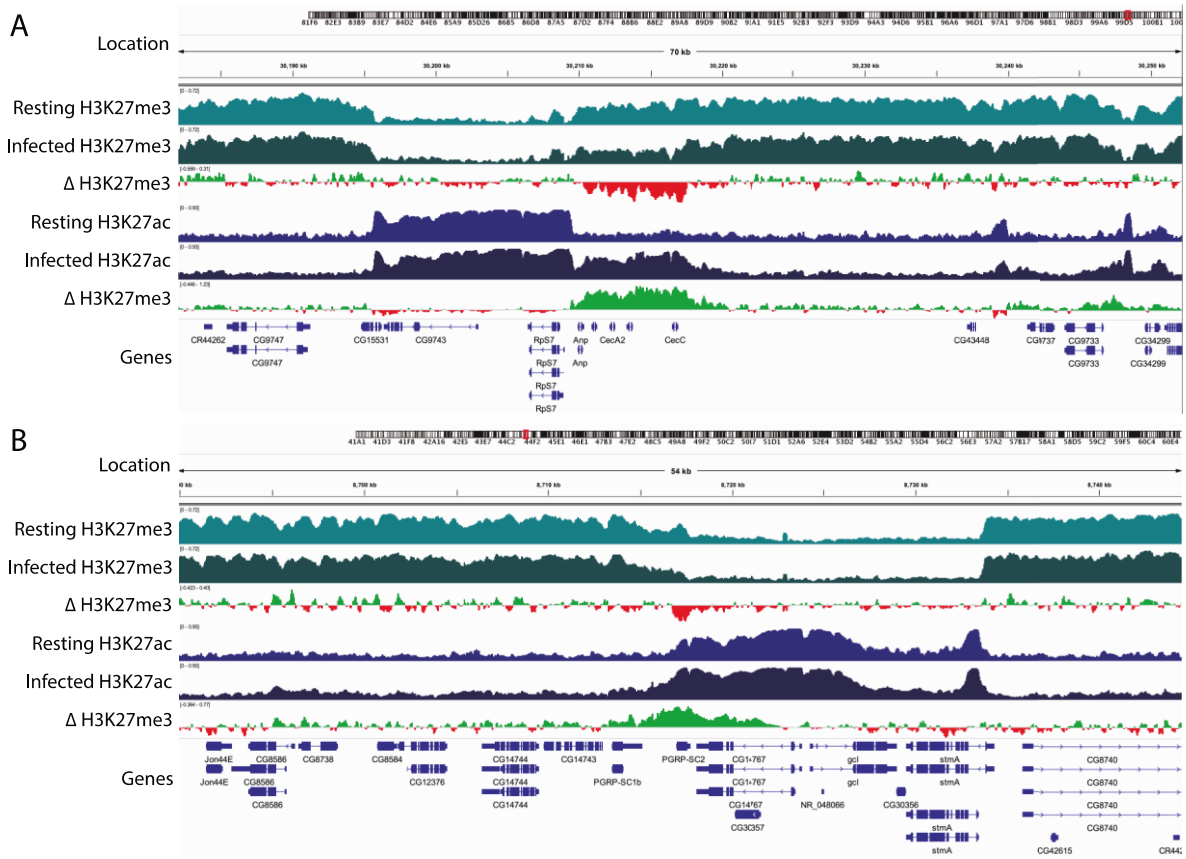

Figure 27: Select immune induced genes are reduced in H3K27me3 and increased in H3K27ac post immune challenge. Enrichments from unchallenged (resting) and 6h post septic injury (infected) plasmatocytes for ChIPs of H3K27me3 and H3K27ac were quantile normalized and averaged. Additionally, signal differences were calculated and colored red for regions with a lower signal in the septic injury and green for those with a higher signal. The resulting tracks and their differences were visualized using IGV on two loci. A: Genome track around the *Cecropin* cluster. B: Genome track around PGRP-SC2.

The top panel shows a genomic region around the *Cecropin* gene cluster, a group of AMPs that are differentially regulated in plasmatocytes at 6h after septic injury (compare Figure 3C and Figure 4). I visualized three data tracks for both H3K27me3 and H3K27ac: First, the quantile normalized signal track from unchallenged plasmatocytes, then the same from 6h post septic injury plasmatocytes, and last the difference of the two tracks. Here H3K27me3 is indeed reduced at a region around these AMPs. From the difference track it is clear that this difference is restricted to a small number of genes in the AMP cluster, and does not globally affect H3K27me3 levels, which would indicate a technical artifact of ChIP-seq or normalization. At the same time,

H3K27ac is strongly increased on the same locus but does not appear changed much at neighboring regions. It is possible, that AMPs might be a special case, both because of their short coding sequences and because of their arrangement in clusters. PGRP-SC2, on the other hand, is a pattern recognition receptor that is induced in plasmatocytes after septic injury (compare Figure 3C and Figure 4) and it is shown as an example of a non-AMP immune gene. As for the *Cecropin* gene cluster, there is a strong difference in H3K27me3 at the PGRP-SC2 locus located in the middle of the shown genomic region, which is more restricted to the gene body compared to the *Cecropin* genes. Another difference to the AMP cluster is the topology of the gain in H3K27ac signal: For PGRP-SC2 this gain appears to be spread to regions further outside the gene body, making the region of H3K27ac gain much larger than that of H3K27me3 loss. These changes are overall restricted to the genes of interest and do not globally affect all genes in the region. Together these two examples demonstrate that after quantile normalization the differences predicted by peak calling are well visible on the level of individual genes.

Whereas this way of data representation is very intuitive, information about replication, which is useful to determine the reproducibility of the effect, is lost. Therefore, I calculated the quantile normalized gene-level H3K27me3 enrichment of each genes that was up-regulated at 6h post septic injury. Of those genes I selected ones that either overlapped with the repressive chromatin state 3, or were H3K27me3 positive according to the H3K27me3 gene state (Figure 12 and Figure 13). I plotted the resulting values as a z-score normalized heatmap (Figure 28). In this heatmap of H3K27me3 differences, columns represent samples, with both ChIP-seq samples from unchallenged plasmatocytes to the left and both post septic injury samples to the right. Further, I grouped genes, which are individual rows, first based on if they either overlapped chromatin state 3 or were H3K27me3 positive or both and then sorted them by hierarchical clustering.

Overall, the two control samples to the left have a lot more blue heatmap tiles than the septic injury samples, indicating an overall higher level of H3K27me3 level on genes in these samples.

In this heatmap, there are very few genes that are only in H3K27me3 positive state, but the treatment well predicts the H3K27me3 enrichment, with both control replicates being higher in H3K27me3 for all genes, while the septic injury ChIP-seq samples are always lower. For the genes that only overlapping chromatin state 3, on the other hand, there are replicates that are lower for H3K27me3 in the control sample than in the septic injury samples. However, it is important to remember that chromatin state assignments are not exclusive, and it is possible that these genes are incorrectly assigned to repressive chromatin states.

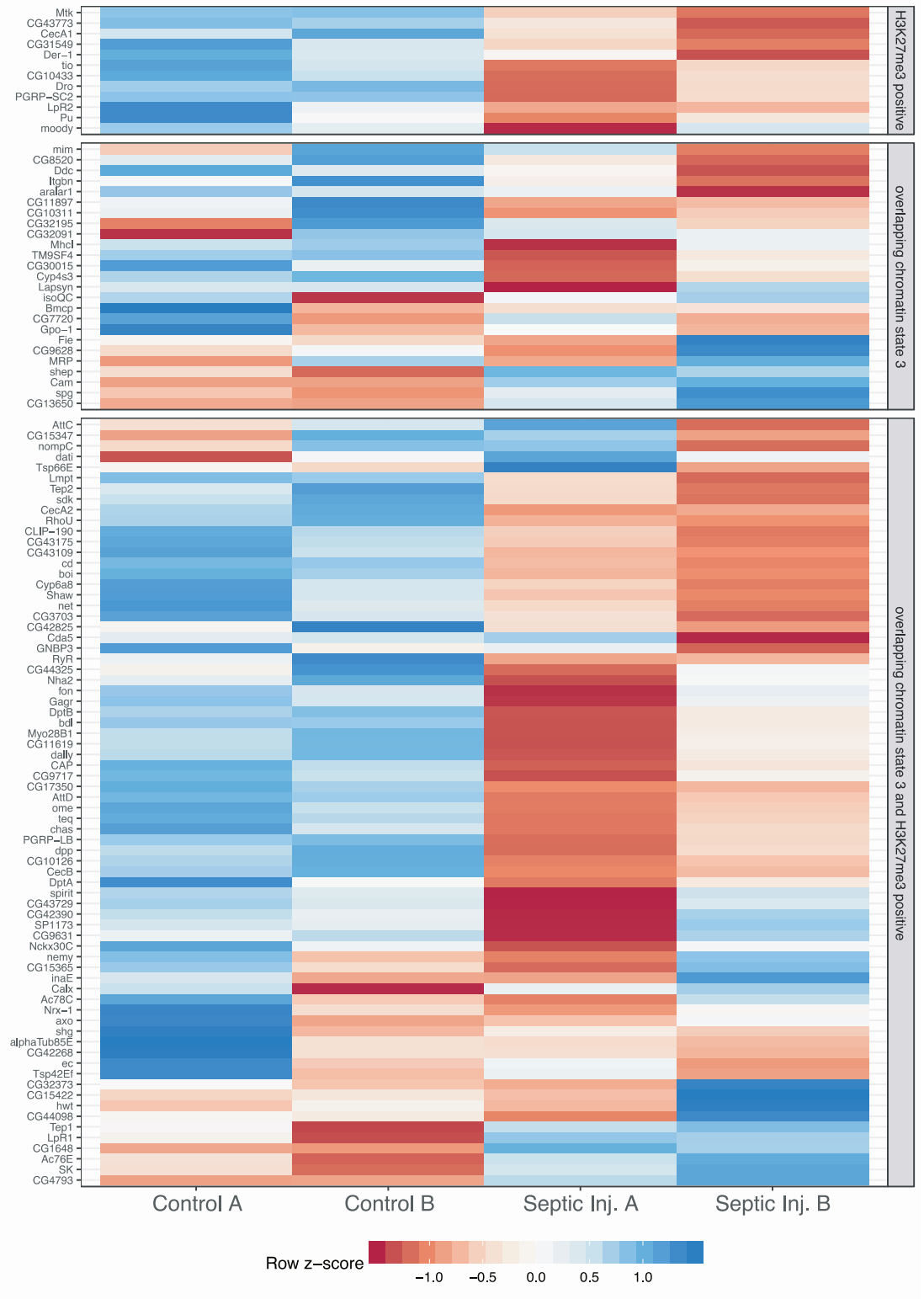

Figure 28: A small number of genes show discordant H3K27me3 regulation at 6h post septic injury. Quantile normalized enrichment of H3K27me3 was calculated on a gene level for both unchallenged (control) replicates and for both 6h post septic injury replicates. Genes that were up-regulated at 6h post septic injury by RNA-seq were filtered for genes that are H3K27me3 positive (higher than the background) or overlapping with chromatin state 3. The gene-level H3K27me3 enrichment signal was z-score transformed and plotted as a heatmap.

For the genes that are both in an H3K27me3 positive gene state and overlapping with the chromatin state 3 the vast majority of genes lose H3K27me3 after septic injury, with at least one of the septic injury samples showing a strong loss of H3K27me3. Specifically, these are the genes in the middle of the heatmap, and they include many well-characterized immune genes including AMPs from *Diptericin*, *Attacin* and *Cecropin* clusters, TEPs and PGRPs.

In summary, this demonstrates that across all H3K27me3 positive immune induced genes, most of them reproducibly loose H3K27me3 after immune challenge, but a small number of genes responds discordantly

Therefore, I characterized population-level histone modifications changes in greater detail. For H3K27me3 I used the same quantile normalized gene-level enrichment applied to each of the H3K27me3 ChIP-seq experiments that were generated for the heatmap. For H3K27ac, some modifications are necessary: As shown earlier, H3K27ac is restricted mostly to regions around the transcriptional start site (TSS, see Figure 10). Therefore, if I calculated the enrichment on a gene level, this enrichment would be a mix of H3K27ac positive region at the start of the gene and H3K27ac negative regions towards the 3' end. This would, therefore, mean that the signal would differ not only based on the height of the signal (which is the intention) but also by the length of the gene (which is not correct). Hence, I restricted the analysis to the TSS of each gene by counting reads for 500 bp downstream of all potential TSS of each gene, calculating their enrichment and for each gene choosing the one TSS that had the highest enrichment (which is, therefore, the most likely site from which a gene product is transcribed).

Using these H3K27me3 and H3K27ac enrichment sets I calculated some population statistics. First, I generated 4 categories by dividing either genes that were in a H3K27me3 positive gene state (Figure 12) or that overlapped with chromatin state 3 (Figure 13) and then further dividing genes into those that were up-regulated at 6h post septic injury in plasmatocytes (later just called up-regulated) and those that were not up-regulated. To identify a confidence interval of H3K27me3 and H3K27ac enrichment for each category I applied bootstrapping[285]: From each group, many random samples are drawn and the mean is calculated each time. By repeating this many times, it is easy to find the interval in which 95% of all calculated means lie, giving a good estimator for the 95% confidence interval of the population mean. In this way, I calculated means and confidence intervals for all samples and groups and plotted them in Figure 29.

For both, H3K27me3 positive genes (left panel) and genes that overlap chromatin state 3 (right panel) among up-regulated genes, H3K27me3 enrichment is reduced in both ChIP-seq replicates from 6h post septic injury plasmatocytes (Figure 29A). At the same time, this loss of H3K27me3 does not appear to be true for the not up-regulated genes. Indeed, if I use the gene-wise mean of

the replicates and calculate p-values applying a paired Wilcoxon signed-rank test I find that for up-regulated and H3K27me3 positive genes $p = 9.56*10^{-10}$ and for up-regulated and overlapping chromatin state 3 genes $p = 1.436*10^{-07}$. Therefore, genes that are targeted by H3K27me3 and are up-regulated at 6h post septic injury in plasmatocytes are reduced in H3K27me3 after the septic injury.

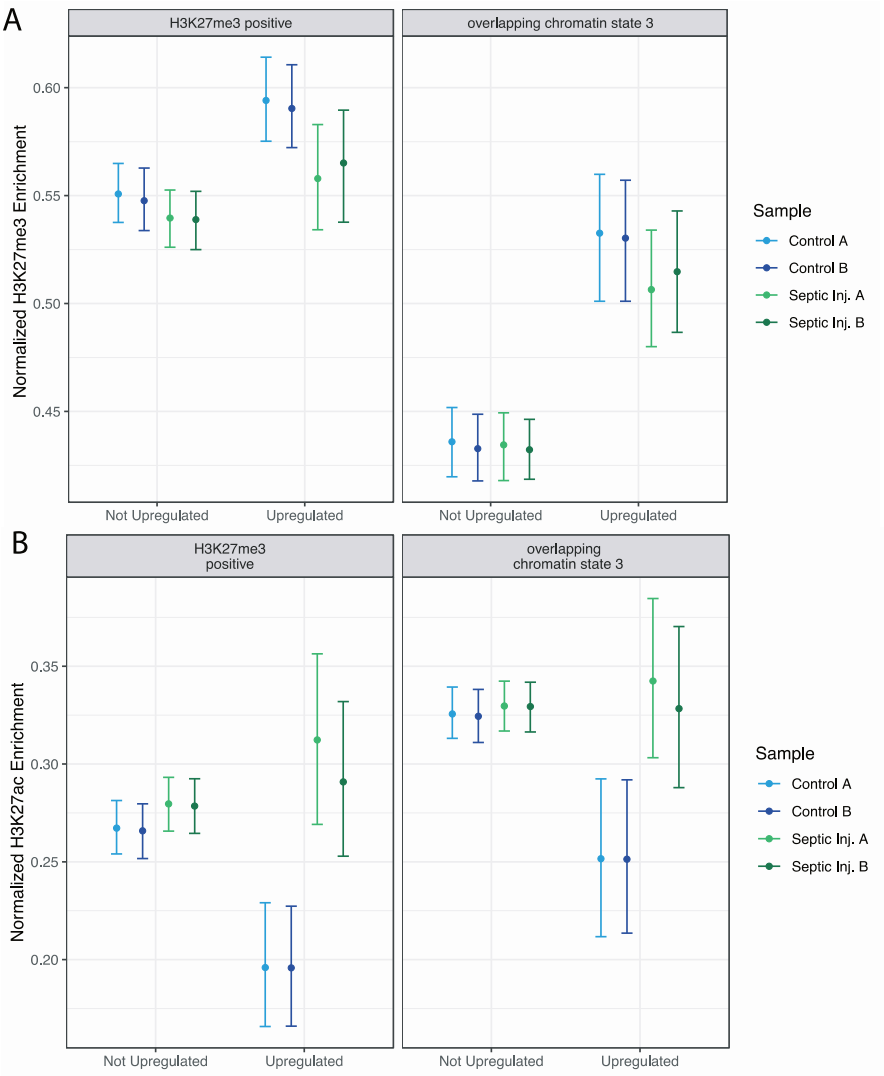

Figure 29: H3K27me3 is reduced and H3K27ac increased across immune induced genes in plasmatocytes after immune challenge. Genes that were either classified as H3K27me3 positive or overlapping with chromatin state 3 were grouped based on their differential expression as up-regulated or not up-regulated. Gene level enrichment (A) or TSS enrichment (B) of each ChIP sample for either H3K27me3 (A) or H3K27ac (B) was calculated and the resulting values were plotted as mean ± bootstrapped 95% confidence interval. Wilcoxon signed-rank test for the means of replicates: H3K27me3 positive up-regulated: (A): $p = 9.56*10^{-10}$; (B): $p = 1.364*10^{-10}$; Overlapping chromatin state 3: (A): $p = 1.436*10^{-07}$; (B): $p = 2.035*10^{-11}$

The reverse seems to be true for H3K27ac at the TSS (Figure 29B). Here genes that are up-regulated at 6h post septic injury have increased levels of H3K27ac at 6h post septic injury. This gain, too, is highly significant and using a paired rank-based Wilcoxon signed-rank test I find that the increase in H3K27ac for up-regulated and H3K27me3 positive genes has $p = 1.364*10^{-10}$ and for up-regulated and overlapping chromatin state 3 genes has $p = 2.035*10^{-11}$. At the same time, the small increase observed for H3K27me3 positive not up-regulated genes only has p = 0.1907. Hence, the same group of genes that was reduced in H3K27me3 after septic injury also has a mean increase in H3K27ac. Together, population statistics, therefore, support the observation, that in plasmatocytes H3K27me3 is lost on immune regulated genes, while at the same time H3K27ac is increased at those genes.

With regard to the mean changes in H3K27me3 (Figure 29A), it might be objected that the changes are only small, and it might be questioned whether these changes affect transcription. However, there is very little data on how H3K27me3 levels affect transcription quantitively. Therefore, it is difficult to say whether the observed changes in H3K27me3 are comparable to the changes that are to be expected based on the transcriptional fold changes. For that reason, I used a linear model to predict the loss in H3K27me3 quantitively. As a basis for the model, I applied my data sets of transcription and histone modification profiles from unchallenged plasmatocytes. I reasoned that, if distinct levels of H3K27me3 at one individual gene were associated with distinct levels of transcriptional output of that gene, this will to some degree translates to all genes, such that H3K27me3 levels will correlate with transcriptional output. This assumption is likely vastly oversimplified, as it does not consider the effects of transcription factors, other histone modifications, or gene intrinsic properties. But even if a model does not fully explain the data variance it could be a good estimator to compare to the magnitude of transcriptional and H3K27me3 changes observed in plasmatocytes after challenge. To build this model I first selected only those genes which are expressed with at least 1 read mapped to them. The read counts from unchallenged plasmatocytes of these selected genes were transformed into tags per million (TPM), which normalizes gene-level read counts for both gene length and sequencing depth. I then extracted 2 groups from these genes: First, all genes that do not overlap with any of the repressive chromatin states 1, 3 or 5, and second, all genes that overlapped with repressive chromatin state 3. For these genes, I then compared the average TPM to the mean H3K27me3 enrichment of ChIP-seq from unchallenged plasmatocytes. These correlations are plotted in Figure 30. The top panel here shows the correlation between H3K27me3 and $\log_2$ TPM for all genes that do not overlap any of the repressive chromatin states, where every dot is an individual

gene and the blue line is a linear model of their relation. Here, genes are expressed at different levels mostly independent of their H3K27me3 levels. Therefore, even though all genes lie very close to the linear least-squares model, it does not appear to predict a lot of the change. Pearson correlation coefficients are r = -0.075, with p = $2.96*10^{-9}$, which means there is only a weak (but significant) anticorrelation between expression strength and H3K27me3 levels in this gene group.

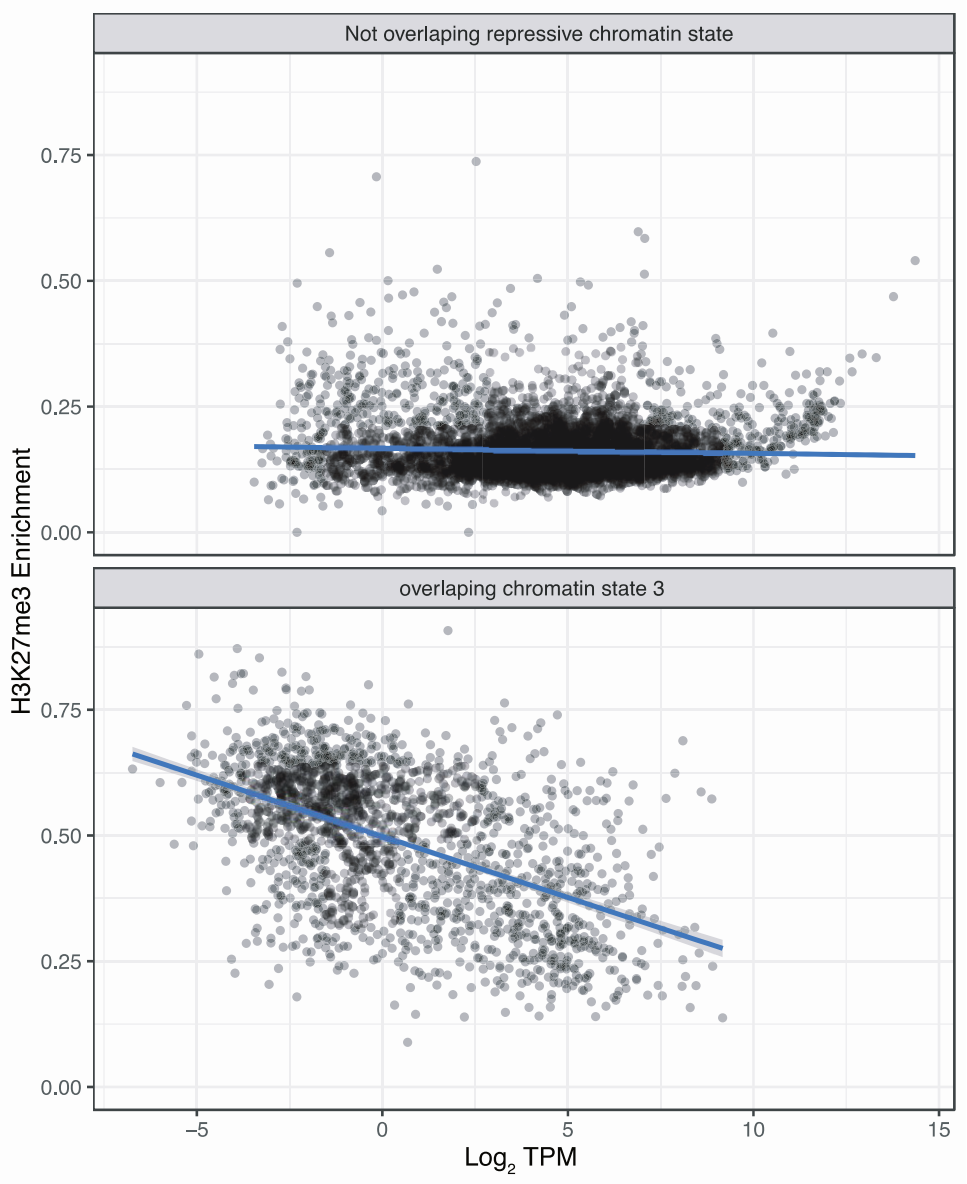

Figure 30: H3K27me3 enrichment and expression strength correlate in individual chromatin states. Gene level expression strength as $\log_2$ TPM compared to H3K27me3 enrichment. Genes were grouped into genes that do not overlap a repressive chromatin state (chromatin states 1, 3 and 5) and those genes that overlap chromatin state 3 and have any RNA-seq reads mapped to them. Blue lines show linear models. Not overlapping repressive chromatin states: r = -0.075, p = $2.96*10^{-9}$, Overlapping chromatin state 3: r = -0.556, p < $2.2*10^{-16}$.

This correlation is much stronger when I analyze those expressed genes which overlap with chromatin state 3 (Figure 30, lower panel). Even though there are much fewer data points in this group, there is a clear anticorrelation observable, which has a correlation coefficient of r = -0.556 with $p < 2.2*10^{-16}$. This Pearson's r of -0.556 demonstrates that there is a lot of variance in the data that is not explained by the model. But as I discussed above, this model is extremely reductionist, and given the strong statistical significance, it may still be useful to explain the trend in data correlation. Therefore, I extracted the underlying least-squares linear model which is:

$$H3K27me3\ Enrichment = -0.028 * log_2(TPM) + 0.505 \quad \textbf{(20)}$$

This model can then be applied to the results of the RNA-seq of the 6h post septic injury plasmatocytes compared to unchallenged plasmatocytes, where H3K27me3 changes can be predicted by calculating

$$\Delta_{H3K27me3\ Enrichment} = -0.028 * log_2 FC \quad \textbf{(21)}$$

By applying this model to predict gene-level changes in H3K27me3 based on the log_2 fold changes from the RNA-seq, the magnitude of observed changes in H3K27me3 can be compared quantitatively to the model predictions.

Hence, I took all genes that were differentially expressed at 6h post septic injury when compared to unchallenged plasmatocytes and calculated their predicted change in H3K27me3 enrichment using (21). Then, I split genes into H3K27me3 positive or negative genes. From these comparisons of predicted and observed and plot them against each other (Figure 31).

In this comparison of predicted and observed H3K27me3 changes RNA-seq log_2 fold changes predict H3K27me3 changes well for H3K27me3 positive genes, but not for H3K27me3 negative genes. The 181 H3K27me3 negative genes (green) are overall very close to zero on both observed and predicted changes. This is well consistent with the earlier observation that H3K27me3 negative genes have, on average, a smaller log_2 fold change (compare Figure 20). Additionally, all genes in this H3K27me3 negative category appear scattered without a clear trend in the observed H3K27me3 changes, which is well reflected in the Pearson correlation coefficient r = 0.054, p = 0.468. Therefore, there seems to be no correlation between the observed H3K27me3 change and the transcriptional changes for H3K27me3 low genes. For the 87 H3K27me3 positive genes (orange) there is a clear positive correlation, such that larger observed differences in H3K27me3 are reflected in larger predicted changes, based on the linear model from (21). Indeed, I find correlation coefficients of r = 0.508 with $p = 5.1*10^{-7}$. Importantly, the magnitude of the change for observed and predicted H3K27me3 changes seem well comparable. Therefore, in

plasmatocytes fold expression changes quantitively predict the loss of H3K27me3 at H3K27me3 positive genes which are up-regulated at 6h post septic injury, but the same is not true for H3K27me3 negative genes.

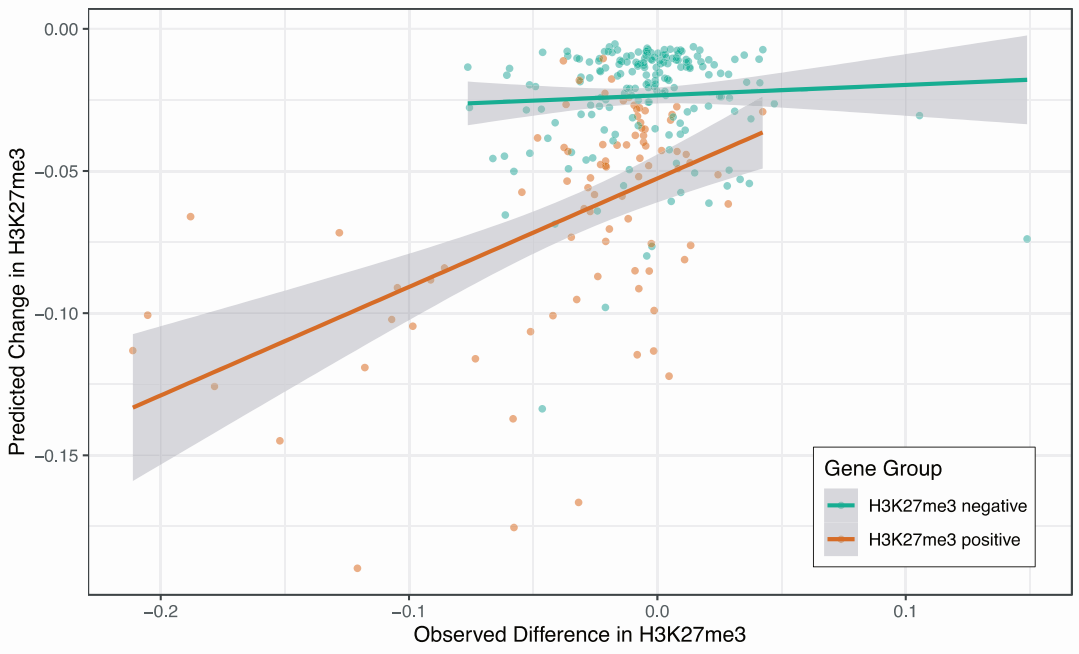

Figure 31: Transcriptional changes post immune challenge predict H3K27me3 loss in repressed genes, but not in active genes. Gene level log-fold changes in plasmatocytes at 6h post septic injury when compared to unchallenged cells were used to predict changes in H3K27me3 enrichment levels using (21). Corresponding changes of observed values were calculated from the differential ChIP-seq of H3K27me3 between 6h post septic injury and control. Genes were grouped by their unchallenged H3K27me3 levels and plotted. Lines show linear models ± 95% confidence interval. H3K27me3 negative: r = 0.054, p = 0.468; H3K27me3 positive: r = 0.508, p = $5.1*10^{-7}$.

In summary, when performing differential ChIP-seq on plasmatocytes comparing 6h post septic injury cells to unchallenged cells, both H3K27me3 and H3K27ac are differentially modified, such that H3K27me3 is lost in a number of regions, while H3K27ac is increased in differentially modified regions. This change is tied to genes that are also differentially expressed after septic injury in plasmatocytes (Figure 24) and for H3K27me3, this change is specific to H3K27me3 positive differentially regulated genes (Figure 22, Figure 23 and Figure 31). In consequence, this raises the question, whether the removal of H3K27me3 is a consequence of gene transcription, or if it is instructive in this process, as I proposed.

5.4 Mosaics of H3K27me3 depletion mutants

So far, I showed that a set of immune induced genes is decorated with the repressive histone modification H3K27me3 and that this mark is removed upon their transcriptional activation following an immune challenge. Therefore, I speculated that the modification of H3K27me3 at these loci is functionally relevant to suppress gene activation in unchallenged plasmatocytes. My approach to address this hypothesis was to genetically remove or reduce the modification H3K27me3 in plasmatocytes and assay transcriptional changes in immune genes. If H3K27me3 is indeed instructive in suppressing immune signaling, which may constantly be activated for example due to the presence of a microflora, I expect a transcriptional induction of immune genes after diminishing the level of H3K27me3 modification in the same way that removal of H3K27me3 releases the silencing of epigenetically silenced developmental genes[89].

Here, I can apply the power of the *Drosophila* model: As I discussed in the introduction, in *Drosophila* I can not only mutate those enzymes (writers) that place a histone modification, but I can also mutate histones directly, something that is impossible in any other metazoan. [85] Because mutations in proteins of the PRC2 which writes H3K27me3 as well as mutations in histone H3 lysine 27 are early larval lethal I generated genetic mosaics of these mutations in order to manipulate plasmatocytes at late larval stages.[40,41] Therefore I want to shortly introduce these systems, starting with the histone mutation system. In *Drosophila,* histone genes are organized in histone gene units (HisGUs), short sequences that contain each a single copy of the core histone genes H2A, H2B, H3, and H4 as well as the linker histone H1. All copies of the HisGU are located in a single locus in the *Drosophila* genome, which is located on chromosome 2L proximal to the centromere and called the histone locus. This arrangement in a single locus allowed researches in previous work to delete the entire region of the histone locus by using the DrosDel system, and the resulting deletion is called ΔHis^{C}.[85] Homozygous ΔHis^{C} animals are arrested early in embryonic development, but this arrest can by rescued by supplying at least 12 wild type HisGU copies in transgene, which allows flies to complete development into adulthood.[93] These transgene copies also provide a useful system to study histone mutants: By supplying 12 copies of HisGU, which carry mutant versions of a histone, the importance of for example one amino acid of a histone can test the. This was done in this thesis, where I used a previously published fly line that carries HisGU copies with a $H3^{K27R}$ mutation, that is histone H3 lysine 27 is mutated to an arginine.[89] Homozygous ΔHis^{C} animals who are rescued using these mutant HisGU-$H3^{K27R}$ copies do however die during early larval development[89] and can therefore not be used here, where animals need to complete larval development. This is where the genetic mosaic system is useful. In principle, this system allows for the generation of animals that carry genetically different cells,

for example both wild type and mutant cells, similar to chimeric animals that have been used in other model organisms.[286] While many techniques have been established to generate such genetic mosaics[287] in this thesis I use exclusively the FLP-FRT mitotic recombination technique.[288,289] Because the way in which I use this system to induce mutants in histones or mutants in PRC2 genes differs slightly I want to discuss them separately, starting with the histone mutant case.

The general strategy for generating histone mutant cells is outlined in Figure 32.

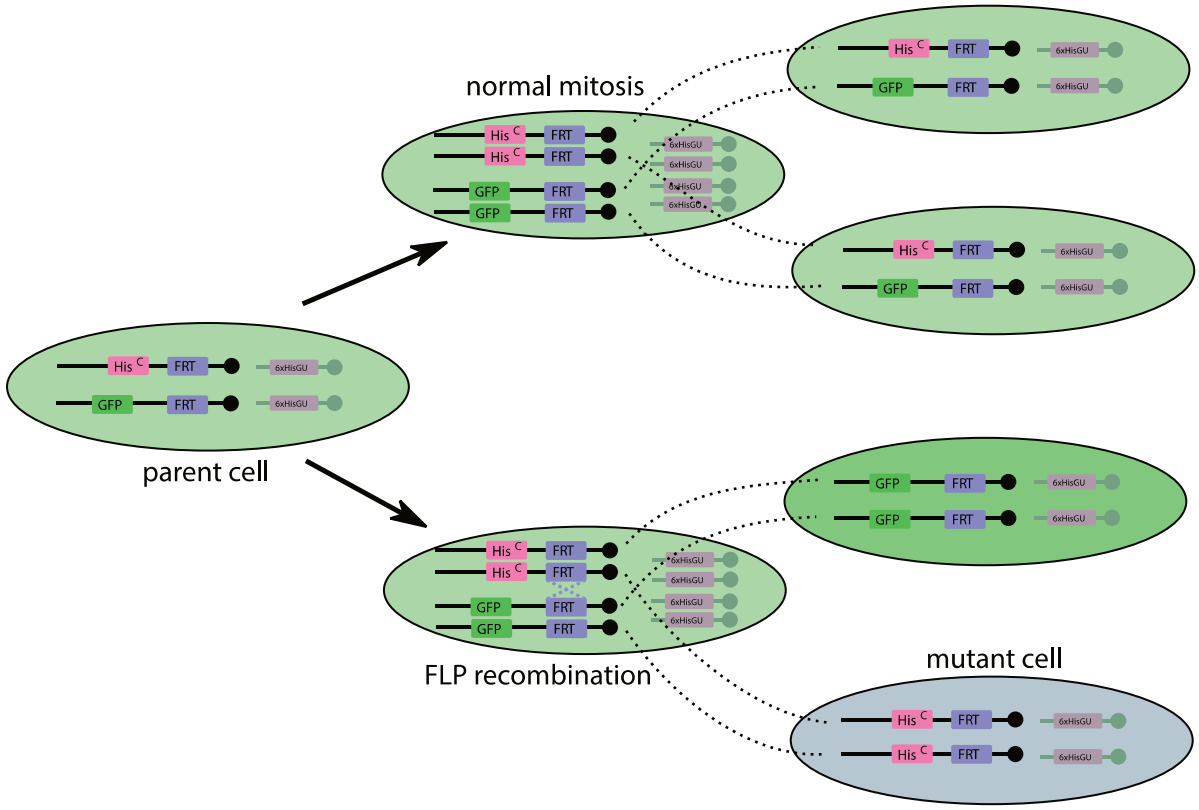

Figure 32: Description of the genetic system used for inducing histone mutant mosaics. Mosaics are induced from a parent cell (left) heterozygous for both ΔHis^C and GFP. Under normal circumstances, when this cell undergoes mitosis, 2 identical daughter cells are produced. If FLP mediated FRT recombination is induced, instead two different cells, each homozygous for either ΔHis^C or GFP are generated. The ΔHis^C is rescued by transgenic histone gene unit copies.

Initially, the parent cell (Figure 32, left) carries FRT recombination sites on both homologous chromosomal arms 2L, distal to which one chromosome carries the histone locus deletion ΔHis^C while the other chromosome is marked by a GFP controlled by the constitutive *Ubiquitin* promotor. This cell is therefore only heterozygous for the ΔHis^C deletion and as such whole animals with this genetic makeup are viable. At the same time, all cells are homozygous carriers

of the transgenic histone gene units (HisGUs), which will later 'rescue' histone locus deficient cells. If FRT mediated recombination is not induced, the cell will undergo regular S-phase DNA replication and the sister chromatids will separate during mitosis to form 2 genetically identical daughter cells (Figure 32, top).

However, if the expression of the recombinase FLP is triggered, which is in this system is accomplished by heat shock mediated induction of a transgene, this recombinase will sporadically cause recombination between the FRT sites of the homologous 2L chromosomal arms in G2 cells (Figure 32, bottom middle). In that case, there is a chance that when sister chromatids separate during mitosis, one of the sister chromatids has exchanged the part distal to the FRT recombination site with that of the homologous chromosome. This can lead to daughter cells that are homozygous for either ΔHis^{C} or GFP (Figure 32 bottom right). Because the ΔHis^{C} homozygous cells lose all GFP expression they can be distinguished from all other cells. At the same time, the transgenic HisGU copies which were present in all parent cells are still present in the mutant ΔHis^{C} cells. But since the cell now lacks any endogenous source of histone, if supplied with mutant HisGU copies (which here are HisGU $H3^{K27R}$) this cell will, therefore, be a histone mutant.

While this exact system has previously been used to generate mosaics in other *Drosophila* organs, for example, wing discs[89], no one had tested whether and at which rate mutant plasmatocytes can be generated in such a fashion. Therefore, in a first test run, I used larvae with the ΔHis^{C} mosaic system (Figure 32) which carried either no transgenic HisGU, 12 wild type HisGU or 12 HisGU $H3^{K27R}$. I heat-shocked these larvae for 1 hour at 1st and 2nd larval stages, and then extracted and FACS sorted plasmatocytes from 3rd instar wandering larvae. I gated the plasmatocytes as described in Figure 2, first for FSC vs. SSC and then for single cells. For the resulting cells, I plotted their GFP levels as a histogram (Figure 33). Similar to previously published data[89,93] only 0.07% of cells are GFP negative for animals that are not rescued by histone gene units, indicating that plasmatocytes which lack histones are unable to proliferate, while both animals with 12 wild type HisGU and 12 HisGU $H3^{K27R}$ produce GFP negative mutant plasmatocytes at 9.59% and 8.51% respectively, demonstrating that both 'rescues' allow plasmatocyte proliferation. Therefore, plasmatocytes from these genetic mosaics can be used for RNA-seq, sorting GFP negative plasmatocytes from 12 HisGU $H3^{K27R}$ mosaics as mutant cells and GFP negative plasmatocytes from 12 wild type HisGU mosaics as matched control cells.

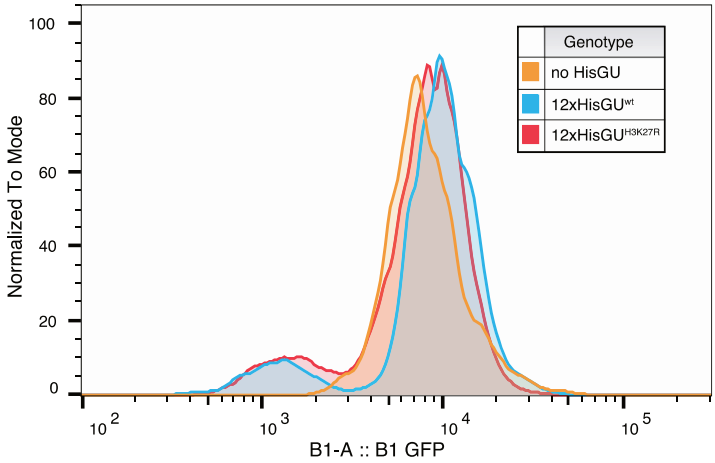

Figure 33: Histone mutant plasmatocytes can be induced as genetic mosaics. Mosaics of plasmatocytes without rescue (orange), with 12x wild-type histone gene unit rescue (blue) or with a 12 histone gene unit $H3^{K27R}$ rescue (red) were FACS sorted. Plasmatocytes were otherwise gated like in Figure 2. GFP negative cells: no His TG: 0.07%; $H3^{wt}$: 9.59%; $H3^{K27R}$: 8.51%.

As I discussed in the introduction, mutating only the histone residue which is modified has shortcomings, since it makes it difficult to distinguish contributions of different histone modifications that occur on the residue (for example H3K27me3 vs. H3K27ac). Therefore, I additionally generated mutants in PRC2, the complex that places H3K27me3 but which has no function in acetylating H3K27.

Hence, I applied a similar mosaic system to generate mutants in the PRC2 components *Enhancer of zeste (E(z))* and *Suppressor of zeste 12 (Su(z)12),* that is none the less different in some key points that I want to mention. *E(z)* and *Su(z)12* are both located on chromosome 3L. Here I used the 2 mutant alleles $E(z)^{731}$ and $Su(z)12^{4}$, which are both premature STOP codon mutants generated by ethyl methanesulfonate mutagenesis[40,290]. For both, *E(z)* and *Su(z)12*, I used previously established chromosomes on which the mutant alleles are located distal to an FRT site, similar to the arrangement in the ΔHis^{C} case (Figure 34). However, on corresponding homologous chromosome I this time used a transgene in which GAL80 is controlled by the ubiquitous *Tubulin* promotor. GAL80 is a transcriptional repressor of the *Saccharomyces cerevisiae* transcriptional activator GAL4, which is frequently used as part of the GAL4/UAS system in *Drosophila*. The function of the GAL80 repressor in this system is to repress the expression of *UAS-GFP* transgene (on the X chromosome) through inhibiting *Tubulin-GAL4* which is ubiquitously expressed from another transgene on the X chromosome. Therefore, the parent cell is here a GFP negative cell that is heterozygous for either $E(z)^{731}$ or $Su(z)12^{4}$ (Figure 34, left).

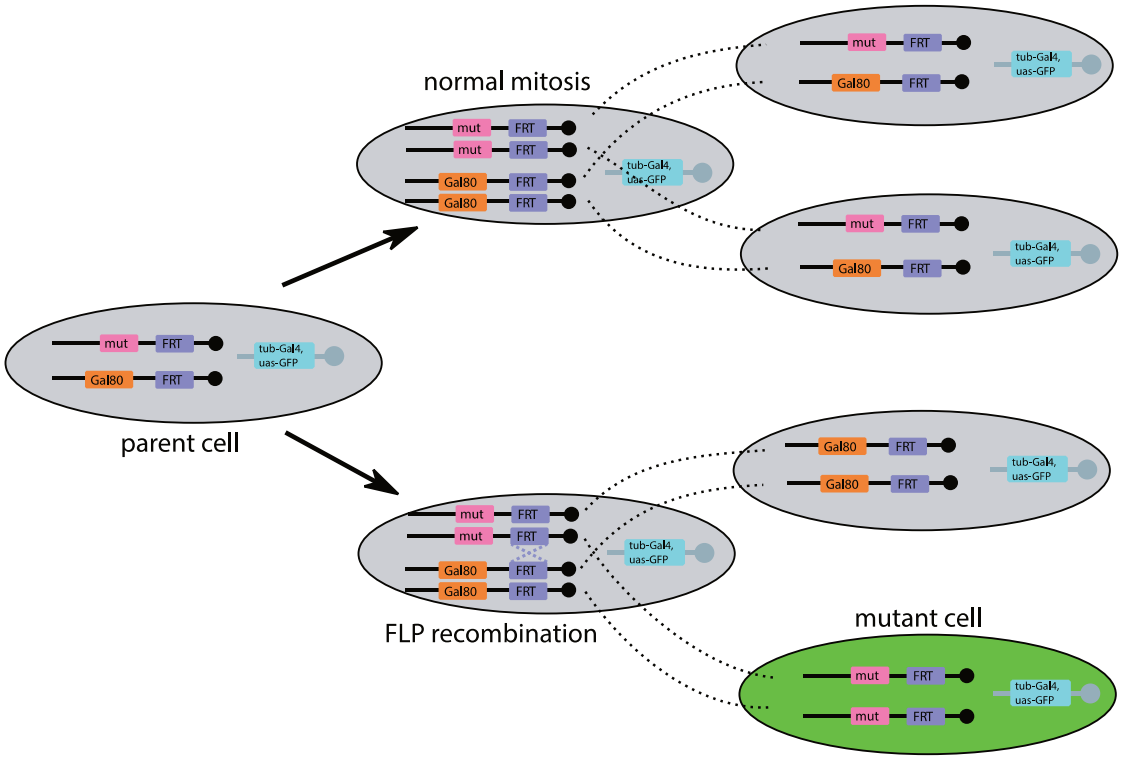

Figure 34: Genetic mosaics used for generating *E(z)* and *Su(z)12* mutants using MARCM. Mosaics are induced from a parent cell (left) heterozygous for either mutant *E(z)* or *Su(z)12* (mut) and a ubiquitously expressed version of the transcriptional repressor Gal80. Under normal circumstances, when this cell undergoes mitosis, 2 identical daughter cells are produced. If FLP mediated FRT recombination is induced, instead 2 different cells, each homozygous for either mutant *E(z)/Su(z)12* (mut) or Gal80 are generated. In cells lacking Gal80, the transcriptional activator Gal4 can trigger GFP transcription.

Now mosaics are induced in just the same way as discussed before. If no FLP is expressed, daughter cells each receive one of the sister chromatids and are heterozygous for both GAL80 and the PRC2 mutation (Figure 34, top). If, however, an FLP mediated FRT recombination occurs, mitotic separation of sister chromatids can lead to 2 daughter cells which are each either homozygous for the Gal80 transgene or for the PRC2 mutation. Since the PRC2 mutants no longer express Gal80, Gal4 can now activate GFP expression in these cells, turning them GFP positive. This system is called MARCM (mosaic analysis with a repressible cell marker).[291] In this case that leads to two key differences: First, the PRC2 mutant cells are GFP positive, while the histone mutants are GFP negative. Second, because the Gal4/uas system strongly amplifies the signal, the brightness of the mutant cells with MARCM is much greater than with the direct GFP construct. This was the main reason why I chose the MARCM system for the analysis of PRC2

mutants since none of the available GFP transgenes from the alternative system generated signals bright enough to detect and separate cells by FACS.

Similar to the protocol I used in the histone mutation mosaics, I took larvae which carried the MARCM transgenes, and either the *E(z)* or the *Su(z)12* mutation, heat-shocked them twice, once each in 1st and 2nd instar larval stages, to induce FLP-FRT mediated recombination, and extracted plasmatocytes from wandering 3rd instar larvae. I first tested the efficiency of the system in plasmatocytes by FACS (Figure 35A). Here I show results from mosaics induced in *E(z)* and *Su(z)12* lines compared to the wild type OrR strain. There is a GFP positive population that is only visible in those genotypes that carry the MARCM system. In the example shown I found about 11.5% mutant cells for the $Su(z)12^{4}$ mutation and 14.5% mutant cells for the $E(z)^{731}$ mutation. Further, I checked the resulting changes in histone modifications, specifically in H3K27me3 and H3K27ac. Therefore in 3 independent experiments, I extracted plasmatocyte mosaics of $E(z)^{731}$ and $Su(z)12^{4}$ from wandering stage larvae, left them to attach to a cell culture dish, fixed and stained them with DAPI and antibodies against either H3K27me3 or H3K27ac and imaged them on a confocal microscope. Representative images of this are shown in Figure 35C-D. Afterwards, I build a custom ImageJ script to identify cells and quantify their signals for GFP, DAPI and the antibody (either H3K27me3 or H3K27ac).

I then normalized the fluorescence signals to the DAPI intensity for each cell, identified GFP positive cells by using a manual cut-off and analyzed the resulting intensities for H3K27me3 and H3K27ac (Figure 35B). Here, H3K27me3 levels are dramatically reduced in both $E(z)^{731}$ and $Su(z)12^{4}$ mutants (Figure 35B, right panel, both $E(z)^{731}$ and $Su(z)12^{4}$: $p < 2.2*10^{-16}$ by Wilcoxon rank-sum test). This is also apparent in the example of $E(z)^{731}$ mutants with H3K27me3 staining (Figure 35C): GFP positive mutant cells (outlined in magenta in each of the channels) are almost depleted for H3K27me3 signal (rightmost panel, red), while GFP negative cells (outlined in yellow) have a homogeneously high signal. Interestingly, $E(z)^{731}$ and $Su(z)12^{4}$ mutants appear to compensate the loss of H3K27me3 with an increase of H3K27ac: For both mutants there is an increase in H3K27ac signal in GFP positive cells (Figure 35B, left panel, both $E(z)^{731}$ and $Su(z)12^{4}$ have $p < 2.2*10^{-16}$ by Wilcoxon rank-sum test). This is also seen in the representative image of $E(z)^{731}$ plasmatocytes (Figure 35D), where GFP positive cells (outlined in magenta) have an elevated H3K27ac antibody staining. While the loss of H3K27me3 after mutation of either *E(z)* or *Su(z)12* is well described and expected by their enzymatic role in this modification[40,41,97], the origin of the increase in H3K27ac is not clear. It is possible that, in the absence of H3K27me3, the histone H3 lysine 27 acetyltransferase *nej* is able to modify nucleosomes more frequently, but on the other hand it is

also possible that due to a loss of PRC mediated repression a number of additional genes become active, which increases the total number of genes targeted by H3K27ac.

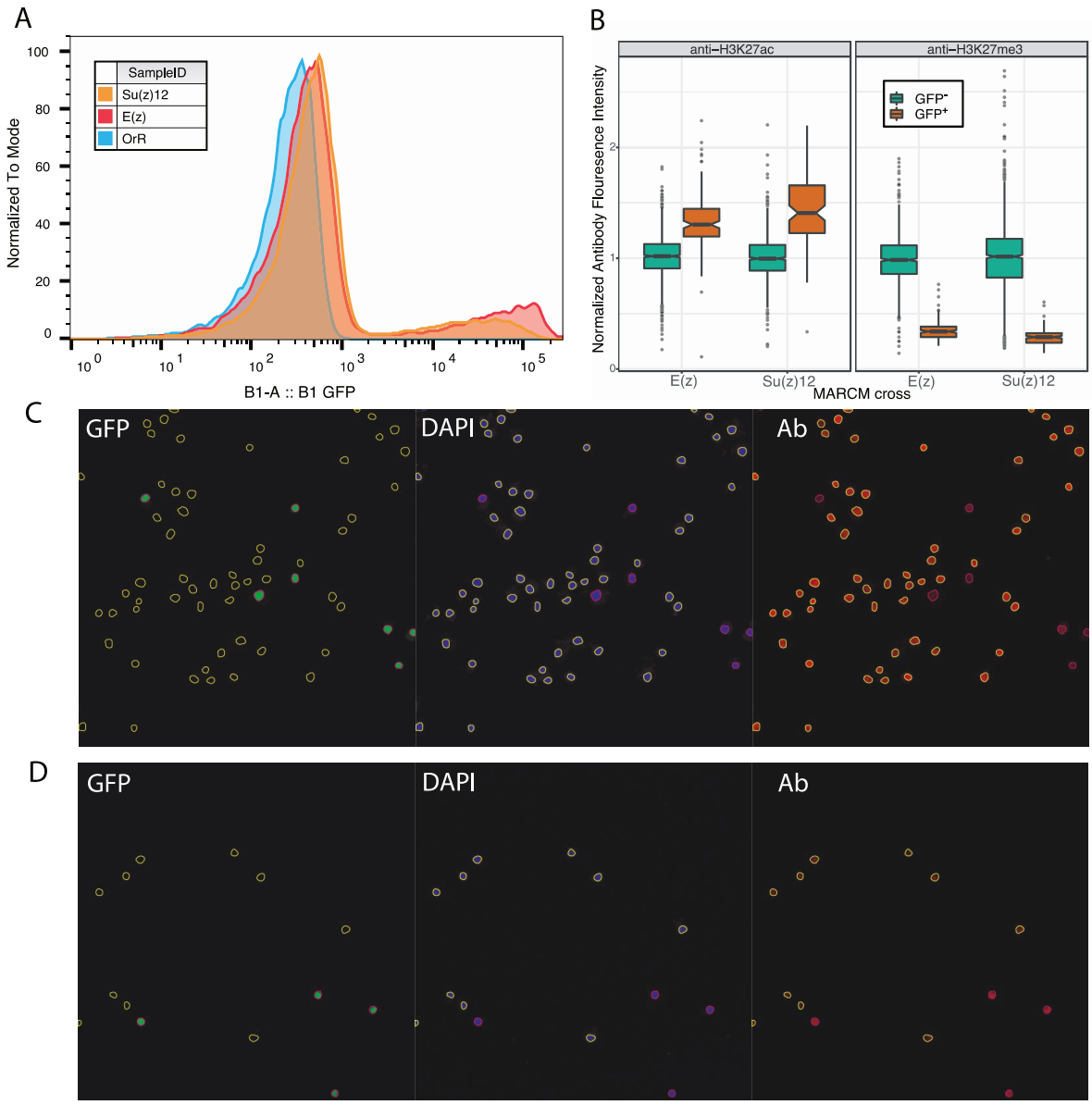

Figure 35: *E(z)* and *Su(z)12* MARCM mosaics can be induced in plasmatocytes and are reduced in H3K27me3. A: Representative FACS plot of MARCM mosaics in plasmatocytes from OrR (blue), *Su(z)12*^{4} (orange) and *E(z)*^{731} (red). Plasmatocytes were otherwise gated as in Figure 2. GFP positive cells: *Su(z)12*: 11.5%; *E(z)*: 14.5%; OrR(wt): 0.08%. B: Quantification of plasmatocyte histone mutant immunostainings. 3 independent experiments like C and D were performed for each genotype and cells were counted from 5-10 separate images. Number of plasmatocytes (as GFP+/GFP-): *E(z)*-H3K27ac: 148/1068; *E(z)*-H3K27me3: 144/940; *Su(z)12*-H3K27ac: 100/1008; *Su(z)12*-H3K27me3: 134/1381. Wilcoxon rank-sum test p-values: all $p < 2.2*10^{-16}$. C-D: Examples of Immunofluorescence of plasmatocyte mosaics of *E(z)*. MARCM mosaics of *Su(z)12*^{4} and *E(z)*^{731} were stain with antibodies against H3K27me3 and H3K27ac. C: Representative image of H3K27me3 staining in *E(z)*^{731}. Green is GFP, blue is DAPI, red is anti-H3K27me3. Yellow outlines are GFP^{-}

cells, magenta outlines are GFP$^+$. D: Representative image of H3K27ac staining in *E(z)^{731}*. Green is GFP, blue is DAPI, red is anti-H3K27ac. Yellow outlines are GFP$^-$ cells, magenta are GFP$^+$.

In summary, I showed that using FLP-FRT based mitotic mosaics I can generate mutant plasmatocytes that are either *H3^{K27R}* or null mutant of *E(z)* or *Su(z)12*. These mutant plasmatocytes can therefore be used to study the role that these factors have in the transcriptional landscape of plasmatocytes.

5.4.1 RNA-SEQ OF PRC2 MUTANT AND *H3^{K27R}* PLASMATOCYTES

Hence, I isolated *H3^{K27R}*, *E(z)^{731}* and *Su(z)12^4* mutant plasmatocytes from the genetic mosaic animals outlined above for RNA-seq by FACS sorting. Additionally, I generated genetically closely matching control samples for each of the mutant cell populations. As a control for *H3^{K27R}* plasmatocytes, I generated mosaic plasmatocytes that are homozygous for ΔHis^C, but are rescued by 12 wild type copies of HisGU instead of 12 HisGU *H3^{K27R}*. For both, *E(z)^{731}* and *Su(z)12^4* I chose to collect GFP negative cells as internal control along with the respective GFP positive mutant cells. These cells are at least heterozygous for *E(z)* or *Su(z)12* and therefore act like wild type.[40,290] After sorting the plasmatocytes as described (Figure 33, Figure 35A) I got 2000 to 7000 mutant cells from each sample (see Table 8). For that reason, I decided to increase the number of replicates in this RNA-seq experiment to account for the larger variance that is to be expected when preparing libraries from smaller cell numbers or lower RNA amounts.[292] In total I generated 5 replicates for both ΔHis^C; 12x HisGU-*H3^{K27R}* and ΔHis^C; 12x HisGU-*H3^{wt}*, and 6 replicates for *E(z)^{731}* and *Su(z)12^4* and their matching GFP negative plasmatocytes. From the RNA isolated from these plasmatocytes libraries were prepared and sequenced (see Table 8). For shortness I will refer to the ΔHis^C; 12x HisGU-*H3^{K27R}*, *E(z)^{731}* and *Su(z)12^4* mutants together as H3K27me3 depletion mutants

After mapping the sequencing data and performing a quality control test, I applied PCA to identify sample groups. Therefore, I determined gene-level read counts and performed PCA on the regularized log-transformed read counts of the 1000 most variable genes across all samples. From this PCA I plotted the first 3 components (Figure 36A). Here, I colored the histone mutants and matching controls in blue, *Su(z)12* mutants and controls in red and *E(z)* mutants and controls in green. Further, the darker colors identify mutant samples, while the lighter colors are the matching controls. In this PCA along the first 2 dimensions the samples vary according to the sample group (ΔHis^C, E(z) or *Su(z)12*) but not based on if the sample was mutant or wild type. Specifically, along PC1, which explains 60% of all sample variance, histone mutant constructs

separate strongly from both *E(z)* and *Su(z)12* samples, indicating a high degree of dissimilarity. It is not quite clear what the origin of these differences are. These samples have very different genetic backgrounds, the samples were generated at different times and library preparation methods differed (see material and methods and Table 8). This should however not prevent a direct comparison of ΔHis^{C} animals rescued with either 12 HisGU^{wt} to 12 HisGU *H3*^{K27R}, since they do not differ along either PC1 or PC2. Interestingly, along PC2, which explains 11.5% of all variance, the samples from *E(z)* and *Su(z)12* genetic backgrounds differ, irrespective of whether they belong to the mutant or control group. In this case, all samples from the 4 conditions (*E(z)* and *Su(z)12* mutant and control) were generated in the same timeframe and libraries were also prepared at the same time with the same library preparation method. Therefore it seems likely that the genetic background is in this case the source of the observed difference. As with the histone mutants, I do want to mention, however, that these potential differences do not affect the analysis here since all samples are only compared in differential expression analysis to their internal control, and not to the other samples. PC3, which explains 9.9% of all variance, is then the component where separation along H3K27me3 depletion is observed: All mutant samples are at the bottom of the plots, and therefore low in PC3, while the matching control samples are at the top and high in PC3. This co-variation along PC3 indicates that there is a common change across the samples that is likely related to the loss of PRC2 related functions in the mutant plasmatocytes.

Therefore, I compared the individual mutant samples with their matching control to identify differentially expressed genes using edgeR. The resulting differences I plotted as volcano plots. For the comparison of ΔHis^{C} 12xHisGU *H3*^{wt} to ΔHis^{C} 12xHisGU *H3*^{K27R} (Figure 36B), there is a far larger number of genes up-regulated than down-regulated (924 up vs. 106 down). This is well consistent with the fact that the ΔHis^{C}; 12xHisGU *H3*^{K27R} mutation causes loss of repression, therefore a large number of genes is up-regulated. Additionally, a set of the most strongly regulated genes is also labeled by name in the volcano plot, and the Hox genes among them are marked in red. Here all Hox genes of the Bithorax complex (*Ubx*, *abd-A*, *Abd-B*) are strongly up-regulated in *H3*^{K27R} mutants, as previous data predicts for the loss of Polycomb silencing.[89] However, the genetic backgrounds of the 12 HisGU *H3*^{K27R} and the 12 HisGU *H3*^{wt} transgenes are not entirely matched, because these previously established constructs were generated in part in different land sites for transgene integration. Therefore, differential expression of individual genes could derive from the *H3*^{K27R} mutation or in some instances, from other differences in genetic background. For the *E(z)*^{731} mutant (Figure 36C) and the *Su(z)12*^{4} mutant (Figure 36D) a larger number of genes is up-regulated than down-regulated in both, *E(z)* mutants (down: 70, up:

355) and *Su(z)12* mutants (down:81, up:733). As for $H3^{K27R}$ mutants, Hox genes are strongly up-regulated in both PRC2 mutants, specifically those from the bithorax complex.

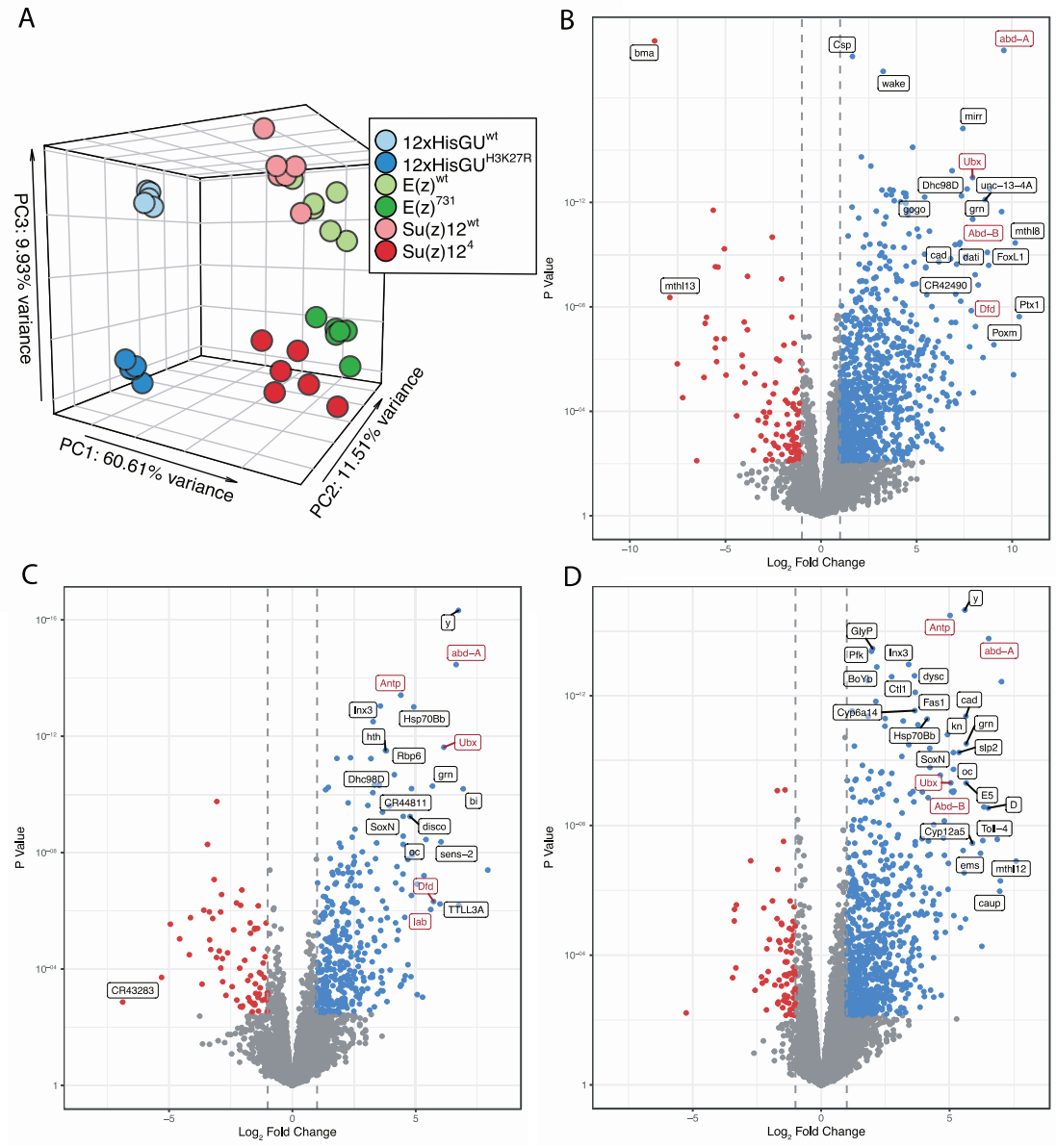

Figure 36: PRC2 and histone $H3^{K27R}$ mutants show concordant transcriptional changes. A: PCA of RNA-seq from sorted mosaics of H3K27me3 depletion mutants from unchallenged plasmatocytes. Mosaics were induced in plasmatocytes and mutant and matching control cells were sorted by FACS. Red: Mosaics in *Su(z)12*, Green: Mosaics in *E(z)*, Blue: Histone cluster mosaics. Light colors indicate wild type plasmatocyte samples, dark colors mutant samples. B-D: Volcano plots of all detected genes for the plasmatocyte mosaic comparisons ΔHis^C; 12xHisGU-$H3^{K27R}$ vs ΔHis^C; 12xHisGU-$H3^{wt}$ (B), $E(z)^{731}$ vs $E(z)^{wt}$ (C) and $Su(z)12^4$ vs $Su(z)12^{wt}$ (D). P-values are plotted log-transformed against the $\log_2$ fold change (LFC). P-values were calculated using a GLM negative binomial model (edgeR) and significant genes were selected with a cut-off false discovery rate corrected p-value = 0.05 and an LFC = 1 (marked by vertical dashed lines). The most outlying genes are marked by name, and Hox genes are marked with red labels.

The difference in the number of up-regulated genes is, however, unexpected, considering that both proteins *E(z)* and *Su(z)12* act as part of the same complex for silencing.[40,41] It is possible that differences in noise levels caused a different number of genes that passed the significance threshold, independent of the true number of differently expressed genes. Additionally, the different up-regulation of *yellow* (y) in $E(z)^{731}$ and $Su(z)12^{4}$, but not $H3^{K27R}$ is striking. This does, however, derive from a previously undescribed uas driven y^{+} expression in the MARCM stock.
Together, this shows that $H3^{K27R}$, $E(z)^{731}$ and $Su(z)12^{4}$ mutants, while their genetic background is diverse, show a component of concordant gene up-regulation that includes many Hox genes.
Since all three mutants, $H3^{K27R}$, $E(z)^{731}$ and $Su(z)12^{4}$, have a large number of genes that are induced I addressed the overlap between those up-regulated genes. It is well characterized that Hox genes are induced after PRC2 loss.[25,89] Therefore I selected the Hox-like homeobox transcription factors according to Burglin and Affolter [293] and calculated from these genes sample wise regularized log-transformed expression values. I plotted these expression values as z-scores in a heatmap (Figure 37A). As described for the RNA-seq after septic injury, I only included genes that had a minimum of 1 read mapped to them per 1 million reads in at least 5 samples. Interestingly, most of the Hox-like homeobox transcription factors do not meet this criterion and out of the 17 genes in this category, 9 are below the read count threshold (*pb, zen, zen2, bcd, ftz, eve, ind, btn* and *ro*). Therefore, these genes appear unaffected by the loss of PcG silencing. The others Hox-like genes are plotted in the heatmap. Here, the genes towards the top of the heatmap, which include the *Bithorax* complex Hox genes and *Antennapedia* (*Antp*) are highly consistent in their up-regulation after H3K27me3 loss. Further, *labial* (*lab*) and *Deformed* (*Dfd*) are slightly less consistent up-regulated, such that they have each 2 outliers in *E(z)* and *Su(z)12* mutants respectively. Interestingly, outliers for these genes are not observed in $H3^{K27R}$ mutants, where they are the most highly differentially expressed across all genes in the heatmap. This is well consistent with the volcano plot (Figure 36), which showed that genes from the $H3^{K27R}$ mutants showed larger fold changes compared to the other mutants. The more pronounced and more consistent activation of repressed genes indicates a more extensive loss of gene repression in the $H3^{K27R}$ mutants when compared to the *E(z)* and *Su(z)12* mutants, though the cause of this remains unclear.
Subsequently, I identified the general overlap of up-regulated genes between the different conditions of H3K27me3 depletion. I selected genes that were up-regulated with at least a 2-fold change and $p < 0.05$, which I then plotted as a Venn diagram (Figure 37B). There is a large overlap, with 137 genes being up-regulated in all conditions, which compares favorably to the 355 total

genes that are up-regulated in *E(z)*^{731}. In fact, of those 355 genes 303 overlap at least with one of the other two conditions, *H3*^{K27R} or *Su(z)12*^{4}. At the same time both, *Su(z)12* and *H3*^{K27R} mutants, have a large number of genes that are only up-regulated in those mutants and in none of the other conditions (399 and 604 respectively). The overall larger number of genes that are up-regulated in H3^{K27R} mutants in comparison to either *E(z)* or *Su(z)12* mutants, will in part explain why these genes are not detected as differentially expressed in the other conditions, but *Su(z)12* mutants only have a slightly smaller number of genes up-regulated than *H3*^{K27R} mutants.

Therefore, it seems likely, that other factors contribute to this non-overlap. Potentially, the differences in genetic background, which I already discussed, contribute to differences in up-regulated genes as well. At the same time, it is possible that *Su(z)12* and *H3*^{K27R} act in part in biologically different ways, for example, that either one of them interacts with other components in an exclusive fashion. In summary, while many genes, and Hox-genes in particular, are concordantly regulated in all H3K27me3 depletion mutants, there are also a number of differently regulated genes. Particularly the *H3*^{K27R} mutant appears to both have the largest number of regulated genes and the strongest induction of Hox-genes.

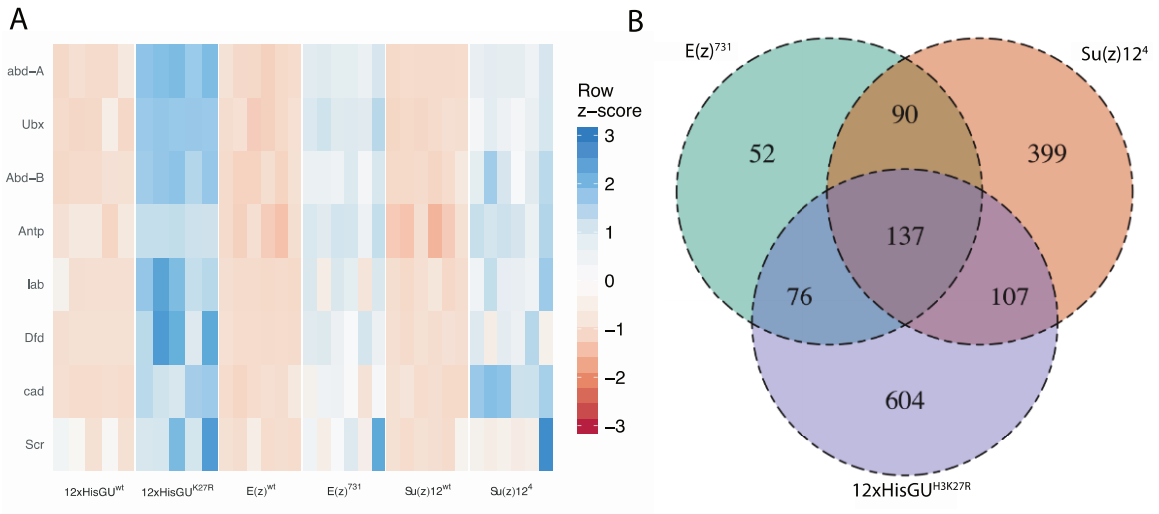

Figure 37: Differentially expressed genes in H3K27me3 depletion mutants. A: Heatmap of Homeobox transcription factors in plasmatocyte mosaics. Regularized logarithmic transformed (DESeq2) gene read counts were normalized by a z-score transformation. Blue indicates higher expression than the mean in that sample, while red indicates lower expression than the mean. B: Venn diagram of up-regulated genes in plasmatocyte mosaics. Genes that were up-regulated when comparing plasmatocyte mosaics of 12xHisGU *H3*^{K27R} vs HisC 12xHisGU *H3*^{wt} (blue), *E(z)*^{731} vs *E(z)*^{wt} (green) and *Su(z)12*^{4} vs *Su(z)12*^{wt} (red) with a false discovery rate corrected p-value < 0.05 and an LFC > 1 were selected. Numbers indicate the number of genes in that overlap.

Further, I determined the type of cellular functions that are disturbed in the H3K27me3 depletion mutants. Therefore, as with the RNA-seq of 6h post septic injury plasmatocytes from chapter 5.1.3, I calculated enriched GO terms: All comparisons of *Su(z)12*^{4}, *E(z)*^{731} and *H3*^{K27R} to their matching non-mutant samples were subjected to the limma goana gene enrichment analysis. I plotted the results, with the size of the circles indicating the effect size of enrichment, the transparency indicates the p-value and the color the directionality of the change (Figure 38).

In addition, I grouped the GO modules into 3 major groups: Developmental terms, metabolic terms and those that fit neither category. First, most GO modules that are enriched significantly are enriched among up-regulated genes, which is well consistent with the observation that there is a far larger number of up-regulated genes than down-regulated genes. Additionally, this indicates that there is no previously annotated biological process that specifically counteracts the loss of H3K27me3. Interestingly, *H3*^{K27R} appears to have the lowest p-values and effect sizes for many of the terms, and developmental terms specifically. At the same time, I showed that the regulation of Hox-like homeobox transcription factors is, if anything, more consistent in *H3*^{K27R} mutants. In general, however, GO annotations in *Drosophila* are much more unreliable than vertebrate model organisms[294], and therefore the completeness and reliability of the terms presented here is, perhaps, debatable. Simultaneously, the samples of *E(z)* and *Su(z)12* mutants are much more similar in their p-values of enriched gene modules to each other than they are similar to *H3*^{K27R} mutants. This observation is both consistent with the level of difference in the PCA (Figure 36) as well as the fact that they are both parts of one protein complex (PRC2) and should, therefore, act together.[40] Importantly, even though the p-values of the enrichments differ strongly between the comparisons, developmental modules are strongly enriched across all genes up-regulated in H3K27me3 depletion mutants.

The reason for the enrichment of metabolic terms is less evident. It is possible, that some level of loss of cell identity causes shifts in metabolism. On the other hand, the terms listed here are either related directly to amino acid metabolism or macromolecule biosynthesis, which includes RNA and protein synthesis. Therefore, these terms could be caused by an overall shift of the cell to increased protein production. This might be caused by a loss of gene repression, which then, in turn, requires an increase in the RNA and protein synthesis machinery. Among the "other" terms, many also relate to development. While not directly containing the term development, genes associated with terms like cell fate specification, regionalization and cell-cell adhesion clearly contribute to it. Additionally, terms related to the regulation of transcription include many developmental genes that are transcription factors. Therefore, it appears that the prime targets of up-regulation in H3K27me3 depletion mutants are both developmental and metabolic genes.

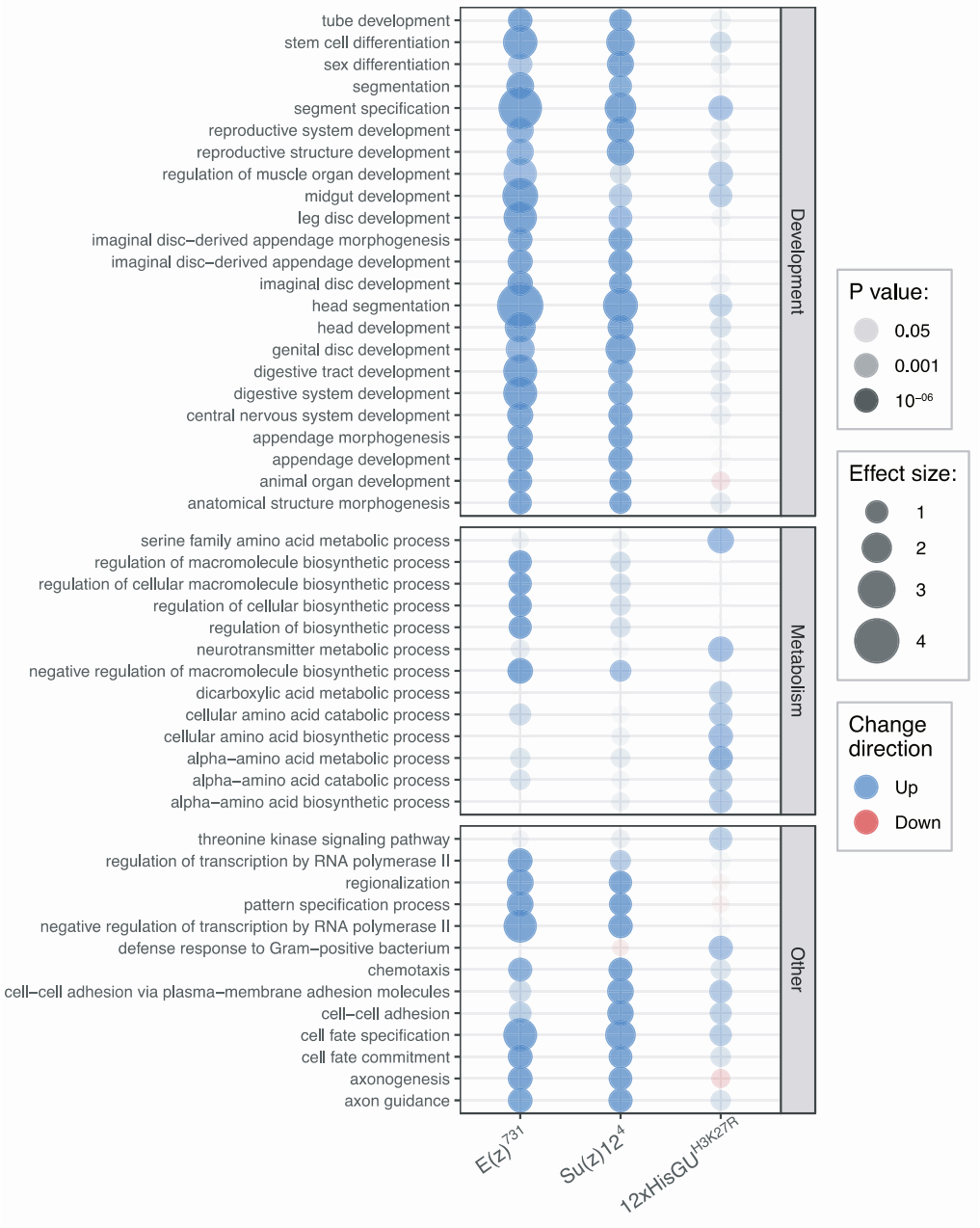

Figure 38: GO-term enrichment in RNA-seq of H3K27me3 depletion mutant plasmatocytes. GO-term enrichment was performed using the goana module implemented in edgeR. Rows are GO modules with the highest significance, while columns are individual comparisons and, where not otherwise noted, the 6h post septic injury plasmatocyte samples were compared to the RNA-seq of control plasmatocytes. The effect size is (b/n)/(B/N) where b is the number of genes that are regulated and carry the GO-term, n is the number of genes regulated in that direction, B is the total number of genes with that GO-term, and N is the total number of genes in the comparison. GO-terms are sorted by manual selection to reflect common groups of gene functions.

In addition to using GO terms to infer biological function, I compared the observed transcriptional changes to both the chromatin states and H3K27me3 gene states I established earlier (chapter 5.2.3). Therefore, I applied the limma fry algorithm to the *E(z)*^{731} vs. *E(z)*^{wt}, *Su(z)12*^{4} vs. *Su(z)12*^{wt} and 12xHisGU *H3*^{K27R} vs. 12xHisGU *H3*^{wt} comparisons using both the gene level overlap with the 7-state chromatin model and the 3-state gene wide H3K27me3 model as gene modules (Figure 39A). Here, as indicated by the blue color, modules are only enriched among up-regulated genes, which is well consistent with the overall small number of down-regulated. Chromatin state 7 is absent from the plot because it was not significantly enriched in any of the comparisons. At the same time, genes overlapping chromatin state 6 are only slightly enriched and only in the *H3*^{K27R} mutants (p = 0.02), which is expected, as chromatin state 6 like chromatin state 7 marks active genes that should be mostly unaffected by the loss of H3K27me3. Unsurprisingly[55], the strongest enrichment, both by effect size and p-value, is observed for genes overlapping chromatin state 5 and H3K27me3 high state genes. These chromatin states correspond to and identify genes strongly targeted by H3K27me3, for example, Hox genes. Hox-like homeobox transcription factors are not expressed in wild type plasmatocytes (Figure 37A) and a subset of them is consistently and strongly induced after H3K27me3 loss, which makes a strong enrichment of the H3K27me3 high state likely.

Further, there is some level of enrichment for both of the enhancer chromatin states that I identified, chromatin states 2 and 4. Interestingly chromatin state 2 is consistently more strongly enriched by effect size and p-value compared to chromatin state 4, which is only enriched in *Su(z)12*^{4} and *H3*^{K27R} mutants. This might be explained by the hypothesis that chromatin state 2 represents a more repressed type of enhancer, while chromatin state 4 are active enhancers. Last, there is an enrichment for the three repressive gene states which are targeted by intermediate levels of H3K27me3, chromatin state 1 and 3 as well as H3K27m3 medium genes. These states are all very similar in their enrichment, both in effect sizes and p-values for each of the comparisons, and they are very strongly enriched. This indicates that H3K27me3 intermediate levels are indeed instructive for silencing in unchallenged plasmatocytes. The lower levels of enrichment of H3K27me3 medium states compared to H3K27me3 high states has 2 confounding factors: In contrast to H3K27me3 high state genes, not all H3K27me3 medium state genes are fully silent in unchallenged cells, and H3K27me3 high state genes respond either very strongly to the loss of H3K27me3 or not at all. Therefore, the list of most highly differentially expressed genes is led by the H3K27me3 high state genes, after which the H3K27me3 medium state genes follow. Hence, the H3K27me3 high state genes will appear to have a higher gene

module enrichment, but this does not mean that the loss of repression at H3K27me3 medium state genes is any less relevant than that at H3K27me3 high state genes.

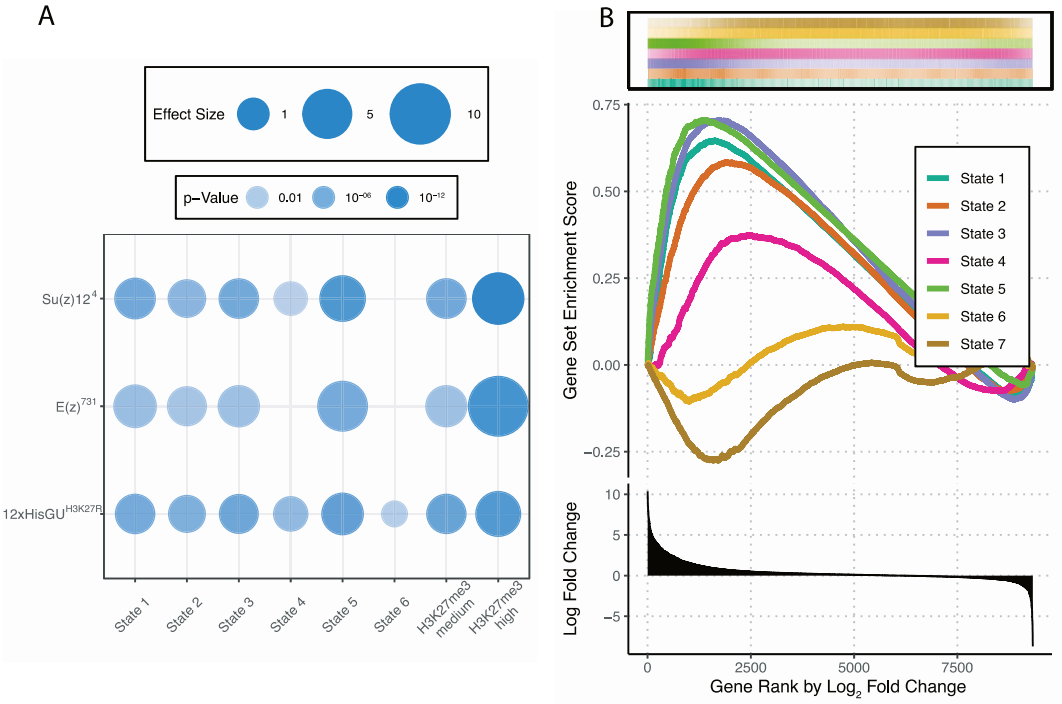

Figure 39: Enrichment of chromatin states among H3K27me3 depletion mutants. A: RNA-seq comparisons of $H3^{K27R}$, *Su(z)12*^{4} and *E(z)*^{731} mutants to their wild type matched samples are analyzed for enrichment of chromatin states and H3K27me3 gene states (Figure 12 and Figure 15) using the fry algorithm. B: GSEA of 12x HisGU $H3^{wt}$ vs. 12x HisGU $H3^{K27R}$ checking for the enrichment of chromatin state overlap. Bottom panels: Log_2 fold change of $H3^{K27R}$ vs. wild type *H3* over gene rank. Top panels: Density of genes positive for the chromatin state. Middle: GSEA score by gene rank. P-values by permutation: Chromatin states 1, 2, 3 and 5: $p < 1*10^{-5}$; Chromatin states 4,6 and 7: n.s.

Hence, I checked if this enrichment patterns – strong enrichment of H3K27me3 medium state genes and H3K27me3 high state genes, with a slightly higher enrichment of the H3K27me3 high states – is also reproduced by an alternative analysis method. Therefore, I applied the gene set enrichment analysis algorithm (GSEA)[280,295] to visualize the enrichment of chromatin states and calculate their p-values (Figure 39B). Here I sorted genes by their log_2 fold changes between 12xHisGU $H3^{K27R}$ and 12xHisGU $H3^{wt}$ (bottom panel), then marked genes by their overlap with either of the 7 chromatin states, whose density is plotted in the top panel. Then enrichment scores were calculated as described in the methods. The results are similar to the fry algorithm enrichment statistics. The enrichment scores of genes overlapping chromatin state 5 rises the

fastest and hits its maximum enrichment score at a very low gene rank. Only a few gene ranks afterwards however, the H3K27me3 intermediate chromatin state 3 also reaches a similar enrichment score. This indicates that both states are similarly de-repressed, however, the extent of the response after the loss of silencing is greater in H3K27me3 high genes in comparison to H3K27me3 intermediate genes. At the same time, the gene module of genes overlapping chromatin state 1, which I appeared to be constitutive heterochromatin, hits its maximum enrichment score around the same gene rank as chromatin state 3, but that enrichment score is lower, indicating a weaker overall enrichment. Further, chromatin state 2 is here enriched among up-regulated genes, perhaps showing that this is a repressed enhancer state that is activated by the loss of H3K27me3. The enrichment of all these 4 chromatin states (1, 2, 3 and 5) is highly significant with $p < 1*10^{-5}$. Last, in GSEA neither of the chromatin state 4, 6 or 7 are significantly enriched, which is in contrast to the fry-algorithm based analysis, where states 4 and 6 were significant. This likely reflects the differences in rotation and permutation-based analysis methods. The p-values of the fry algorithm are however reflected in the order of the maximum enrichment scores produced by the GSEA: chromatin states 4 has the highest enrichment scores among the active chromatin states, while chromatin states 7, which was not significantly enriched using the fry algorithm either, has the lowest enrichment score among all states. In summary, this demonstrates that all repressive states rely to some extent on H3K27me3 mediated silencing, as genes that overlap these repressive states are induced in H3K27me3 depletion mutants. This specifically includes the newly identified H3K27me3 medium state and chromatin state 3, where genes overlapping these states show a highly significant enrichment among up-regulated genes.

Since there is strong enrichment of H3K27me3 high and medium states in plasmatocytes after H3K27me3 loss, this makes it clear that both chromatin states are actively silenced by H3K27me3, irrespective of the level of the modification. Since I showed earlier that immune genes are targeted by H3K27me3, I then characterized how loss of H3K27me3 affects these genes specifically. Therefore, I applied the same analysis that I used above to check for chromatin states now using immune induced gene sets. I defined the set of immune induced genes as all genes that were up-regulated at either 3h, 6h or 18h after septic injury with $p < 0.05$ and fold change > 2. Using these gene sets I applied the fry algorithm as above using the $E(z)^{731}$, $Su(z)12^{4}$ and $H3^{K27R}$ comparisons (Figure 40A). All H3K27me3 depletion mutants are enriched for genes up-regulated 6h post septic injury genes, but this is not the case for other time points after septic injury. Samples from $Su(z)12^{4}$ are not enriched for either the 18h or the 3h post septic injury genes set, while genes up-regulated in $E(z)^{731}$ mutants are not enriched for the 3h post septic injury. At the

same time, $H3^{K27R}$ mutants show the strongest enrichment for 6h and 18h immune induced genes by p-value ($2*10^{-9}$ and $8*10^{-9}$ respectively). This is likely explained by the observation that the $H3^{K27R}$ mutation induced the strongest loss of repression among all the H3K27me3 depletion mutants.

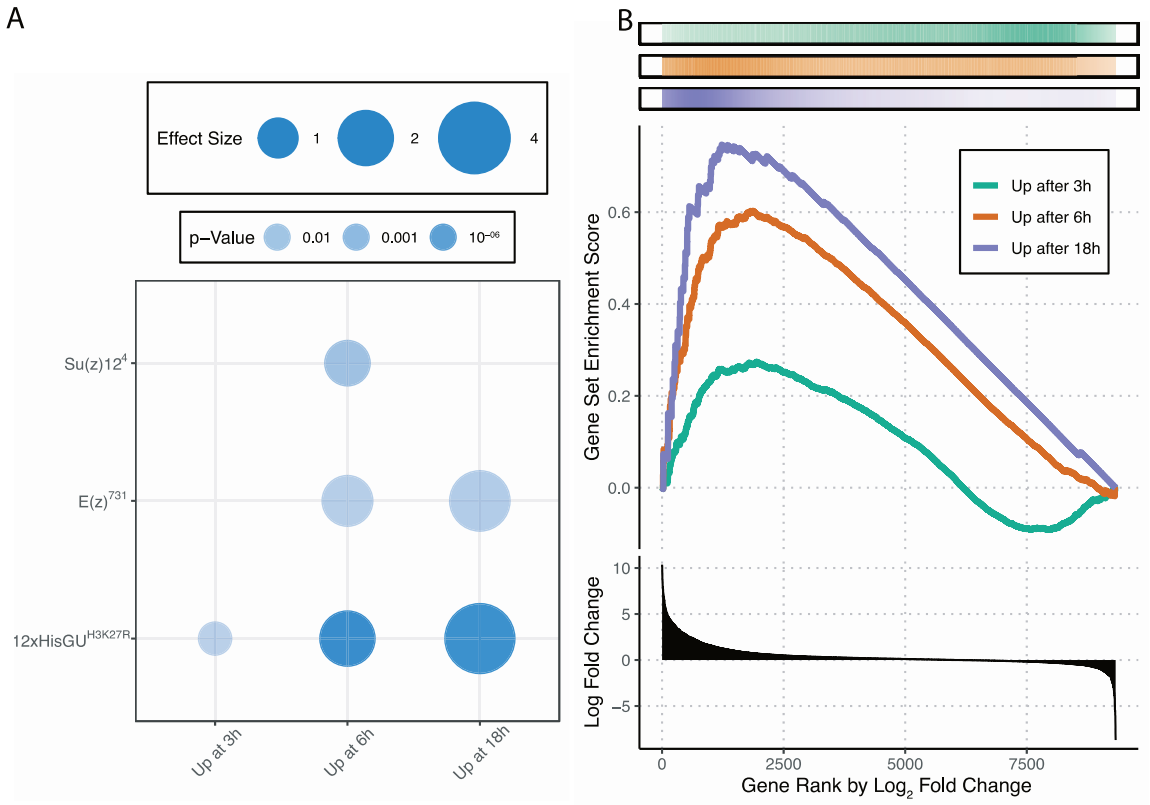

Figure 40: Plasmatocyte immune regulated genes are induced after the loss of H3K27me3. A: RNA-seq comparisons of $H3^{K27R}$, $Su(z)12^{4}$ and $E(z)^{731}$ mutants to their wild type matched samples are analyzed for enrichment of genes up-regulated in plasmatocytes after 3h, 6h or 18h post septic injury using the fry algorithm. B: GSEA of 12x HisGU $H3^{wt}$ vs. 12x HisGU $H3^{K27R}$ checking for the enrichment of genes induced in plasmatocytes after septic injury. Bottom panel: Log2 fold change of 12x HisGU $H3^{wt}$ vs. 12x HisGU $H3^{K27R}$ over gene rank. Top panels: Density of genes positive for the gene module. Middle: GSEA score by gene rank. P-values by permutation: 6h and 18h post septic injury: $p < 1*10^{-5}$; 3h post septic injury: n.s.

Therefore, I applied GSEA using the set of immune induced genes and the 12xHisGU $H3^{K27R}$ over 12xHisGU $H3^{wt}$ comparison (Figure 40B). Here, there is strong enrichment of genes that are up-regulated at 6h or 18h post septic injury (both $p < 1*10^{-5}$), with this time the 18h post septic injury gene set having stronger enrichment. In this analysis method, 3h post septic injury up-regulated genes are not significantly enriched. However, it is important to consider that in the 3h post septic injury-induced genes only a small fraction of them were H3K27me3 positive (Figure 19). Therefore it is likely that perturbance of H3K27me3 levels does not affect genes from this set to the same

extent as those from 6h post septic injury. Together this shows that depletion of H3K27me3 leads to an induction of immune genes, as H3K27me3 depletion mutants are enriched for genes induced in plasmatocytes after septic injury.

The enrichment of immune induced genes in all of the H3K27me3 depletion mutants clearly indicates that those immune genes are generally regulated by H3K27me3. To characterize to what extent the same genes are differentially expressed across all H3K27me3 depletion mutants, and what fraction are affected at all I took the gene set induced in *E(z)*, *Su(z)12* and $H3^{K27R}$ mutants (Figure 37B) and additionally overlaid all genes that were up-regulated at 6h post septic injury and in a H3K27me3 positive chromatin state (Figure 28). I then visualized the overlap between these gene sets as a Venn diagram (Figure 41). Only about half of all genes that are up-regulated at 6h post septic injury and are H3K27me3 positive (green circle) overlap with any other H3K27me3 depletion mutants, and most of them only with the $H3^{K27R}$ mutant. In fact, only 3 of the 6h post septic injury overlaps with all 3 of the H3K27me3 depletion mutants.

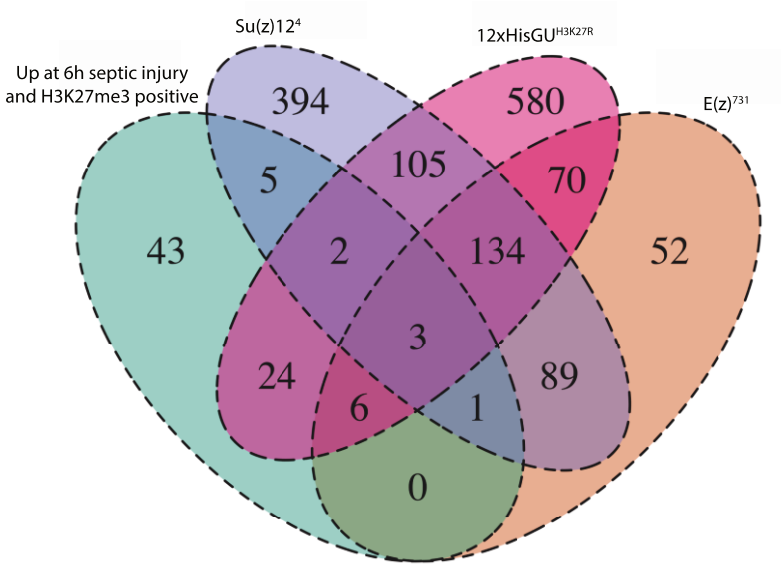

Figure 41: Venn diagram of H3K27me3 depletion mutants and immune induced H3K27me3 positive genes. Venn diagram of up-regulated genes in H3K27me3 depletion plasmatocyte mosaics compared to immune induced genes. Genes were selected for those which were up-regulated when comparing plasmatocyte mosaics of 12x HisGU $H3^{wt}$ vs. 12x HisGU $H3^{K27R}$ (magenta), $E(z)^{731}$ vs $E(z)^{wt}$ (orange) and $Su(z)12^{4}$ vs $Su(z)12^{wt}$ (purple) with a false discovery rate corrected p-value < 0.05 and an LFC > 1. Additionally, genes that were up-regulated with p-value < 0.05 and an LFC > 1 at 6h post septic injury and were in an H3K27me3 positive gene state were selected (green). Numbers indicate the number of genes in that overlap.

This raises the question, why only a fraction of H3K27me3 positive immune genes is affected by the loss of H3K27me3. First, it appears that the number of overlapping genes correlates with the total number of differentially expressed genes. It is possible, that the immune genes are "drowned out" by a large number of other induced genes, that is, the conservative false discovery correction will cut these genes of as potential noise if a large number of other genes is changed more strongly. Second, it is possible that the immune genes only sporadically activate after the loss of H3K27me3 in a fashion that is unrelated to the exact mutation. This would cause higher variability between replicates, and in turn, this would result in weaker phenotypes of H3K27me3 loss, like those observed in *E(z)* and *Su(z)12* mutants, finding fewer significant immune genes. Therefore, it seems likely that some variance in immune signaling and not differences in H3K27me3 regulation cause the differences in up-regulated genes.

Further, I identified how genes that were differentially modified by H3K27me3 and HK37ac after septic injury respond to the loss of H3K27me3 by mutation of writer enzymes or histones. The advantage of comparing the H3K27me3 depletion mutants to differentially modified genes over differentially expressed genes is, that these factors are directly linked: Genes that are differentially modified are more likely to be also regulated by that modification, if such a regulation exists at all, while differentially expressed genes could be differentially expressed for a number of other reasons that are unrelated to H3K27me3.

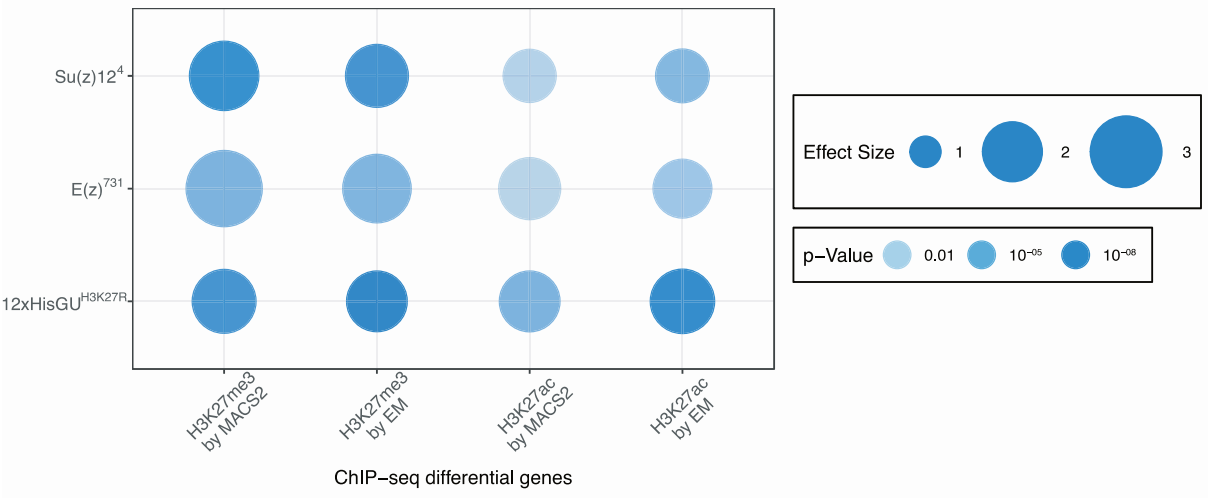

Figure 42: Differentially modified genes and H3K27me3 depletion mutants. RNA-seq comparisons of *H3^{K27R}*, *Su(z)12^{4}* and *E(z)^{731}* mutants to their wild type matched samples are analyzed for enrichment of genes, which were identified to be differentially modified by H3K27me3 (Figure 22) or H3K27ac (Figure 23). Gene groups were identified by applying MACS2 or EM to the differential ChIP samples and finding overlapping genes. Enrichment was tested using the fry algorithm

Therefore, I applied the limma fry algorithm to the set of differentially modified genes that at 6h post septic injury were either increased in H3K27me3 or reduced in H3K27ac, as determined by MACS2 and EM (Figure 22 and Figure 23), using the $E(z)^{731}$, $Su(z)12^{4}$ and $H3^{K27R}$ comparisons and plotted them (Figure 42). As for the enrichment of chromatin states and differentially expressed genes, modules are only enriched among up-regulated genes. Genes that are reduced in H3K27me3 in plasmatocytes after immune challenge according to MACS2 or EM (Figure 42, left columns) are enriched, irrespective of the method used for detecting differential histone modification. Comparing the different mutants with reduced H3K27me3 levels shows that $H3^{K27R}$ produces the strongest enrichment by p-value, while *Su(z)12* mutants are slightly less enriched and *E(z)* mutants are lowest. For genes increased in H3K27ac (Figure 42, right columns) there is a similar pattern, though here the difference between *Su(z)12* mutants and *E(z)* mutants is less pronounced. $H3^{K27R}$ having the strongest enrichment in differentially modified genes is consistent with the overall more prominent loss of repression in the $H3^{K27R}$ mutants, when compared to the PRC2 mutants. But overall, for all mutants, it is apparent that those genes which lose H3K27me3 after septic injury (or gain H3K27ac) are also up-regulated when H3K27me3 levels are reduced genetically. This observation confirms the notion that H3K27me3 is instructive in silencing these genes in unchallenged plasmatocytes.

In summary, I showed that genetic mosaics can be used to generate plasmatocytes that lack the repressive H3K27me3 modification. For this, I used mutants in components of the H3K27me3 writer complex PRC2, specifically in *E(z)* and *Su(z)12*, or a direct histone mutation changing histone *H3* into $H3^{K27R}$. In these mutants, a set of genes that were repressed by H3K27me3 in the H3K27me3 high or H3K27me3 medium states is released from repression. Interestingly, this up-regulation of silenced genes was most pronounced in $H3^{K27R}$ mutants, but it is not clear at this time, whether this is due to differences in the persistence of H3K27me3 mediated repression after mutant clones are induced. Further, I showed that in either of these conditions that deplete the H3K27me3 modification, there is an ectopic up-regulation of immune genes, indicating that H3K27me3 is indeed instructive in repressing immune genes in unchallenged plasmatocytes.

6 DISCUSSION

Histone modifications have been described as important regulators of transcription in *Drosophila*.[14] In this context, the H3K27me3 histone mark and the Polycomb group (PcG) proteins which are responsible for this modification have been extensively studied for their function in development.[113] Here in this work, I expanded this understanding by demonstrating a role for H3K27me3 and PcG proteins in regulating immune genes in Drosophila plasmatocytes. First, I characterized the immune transcriptome and the histone modification landscape of plasmatocytes. I then showed that H3K27me3 is associated with immune genes in a newly identified chromatin state which is distinct from 'classical' Polycomb chromatin found at epigenetically silenced developmental genes. I could demonstrate that the H3K27me3 histone mark is lost concomitant with the activation of immune induced genes and that in resting state plasmatocytes this modification is functionally relevant to prevent erratic activation of immune genes in the absence of a pathogenic insult. In the following I want to briefly discuss the implications that these observations have and what models might be used to explain them.

6.1 Plasmatocytes as a model system for dynamic gene regulation

There is a large body of work demonstrating that *Drosophila* plasmatocytes can respond to a wide range of signals, including JAK-STAT signaling[296], hedgehog signaling[297], Notch signaling[126], TNF signaling[197] and Toll/Imd signalling.[192,298,299] To some, if not all, of these signaling modules plasmatocytes have to respond in a circumstantial fashion, which means not according to a predetermined developmental program but as a response to an stochastic event such as an immune challenge. As a consequence, plasmatocytes display a high degree of transcriptional plasticity, similar to vertebrate immune cells like monocytes.[300,301] At the level of individual genes this plasticity translates into a large number of genes that need to remain inactive for an 'undefined' period of time before they are activated. Clearly, activation of sequence specific transcription factors through the various signaling modules is critical for such dynamic transcriptional responses. On the chromatin level however, this raises two questions. First, if transcription factor based activation is entirely absent for most of the time, how do inactive genes escape permanent epigenetic silencing? Or alternatively, if there is constant low level input from the respective signaling modules, how is aberrant activation of supposedly inactive genes prevented?

Here, I used the immune activation of plasmatocytes by gram-positive and gram-negative bacteria, which signal through the Toll and Imd signaling pathways repectively.[232] Such signaling

will activate NF-κB like transcription factors, thereby promoting the transcription of immune genes.[302,303] I was therefore interested in how the histone modification landscape in *Drosophila* plasmatocytes would contribute to this regulation.

In a first step I therefore characterized the plasmatocyte transcriptome and the expression changes after NF-κB signaling. Prior genome-wide studies on transcriptomes of plasmatocytes used microarray based system[141,192,251], whereas I used RNA-seq. In contrast to microarrays, RNA-seq allows me to accurately determine the absolute transcriptome, in this case of plasmatocytes.[304] Furthermore, prior data on immune induced transcriptome changes used a hemocyte like cell line instead of primary cells, they analyzed only a single time point after immune challenge, and only used gram-negatives for immune activation.[192] Hence, here I can show for the first time the full immune response of plasmatocytes to bacteria, including their temporal dynamics. I can demonstrate, that a number of immune genes is induced by the immune signaling, and that this response is comprised of two separate components, an early response that encompasses a large number of genes, and a late response, which includes many AMPs.

Then I characterized the histone modification landscape of *Drosophila* plasmatocytes with the aim to determine the chromatin state of inactive but inducible (immune-) genes. This is to my knowledge the first ChIP-seq of primary *Drosophila* immune cells and may be the largest ChIP-seq experiment of homogeneous *Drosophila* primary samples today. Other experiments that can be found on Gene Expression Omnibus (GEO)[305] or modEncode[306] include mostly whole animals of various developmental stages, or cell lines. There are however a few ChIP experiments performed on wing discs, adult heads or ovaries. While these are definitely more specific than performing ChIP on whole animals, there still remains a well described heterogeneity in each of these samples: Wing discs are divided in multiple zones that are specific in their gene expression and developmental fate[307], and so are ovaries, which contain among others follicle cells, nurse cells and oocytes[308], and while heads are enriched for neurons they still contain a variety of other cells. Altogether, these tissues are best studied for their developmentally regulated changes in gene expression and do not present a very good model for stochastic gene activation. At the same time however, it could be argued that cell lines are exactly such a model system that is additionally composed of a pure cell population. And while cell lines may indeed be a good model for the respective tissue that they are associated with[309], they definitely have some shortcomings: First, they are not primary, and the process of transformation that allowed them to grow in culture changed their transcriptome, which often specifically involves epigenetic changes in histone modifications, sometimes by mutation in the histone modification

machinery[310-312] Therefore it is essentially impossible to determine to what extend they reflect the situation observed in untransformed cells. And second, *Drosophila* cell lines are genomically highly unstable to a point where even cells in one culture will vary drastically in karyotype[313], and, therefore, cell to cell variability in detecting DNA associated factors by high throughput sequencing have to be considered with care.

For these reasons I think that the ChIP-seq profiles of primary plasmatocytes that I present here in this thesis are a valuable data set that can help to understand how histone modifications are distributed in individual cells in *Drosophila* at a level that will be difficult to improve on.

This assertion however comes with a caveat: It requires the plasmatocytes that I isolate to truly be a homogeneous cell type, which additionally is stable across the manipulations that I perform. While plasmatocytes have been shown to be derived from several hematopoietic origins, namely embryonic plasmatocytes[152,154], lymph gland plasmatocytes[156,158], and plasmatocytes derived from larval hematopoietic hubs[162,163], so far, no classification of plasmatocytes into further subgroups could be established, neither by function nor by transcriptional patterns. It is a valid concern that plasmatocyte population may change as a response to infection and therefore confound the transcriptional changes. Anderl, et al. [151] for instance have demonstrated that after wasp parasitization of *Drosophila* larvae emergency hematopoiesis releases additional hemocytes into circulation. They do however demonstrate that in their model lamellocytes are also released, something I never observed in the septic injury experiments in this thesis. Further, for all RNA-seq experiments with septic injury, RNA yields were determined for each sample (see Table 8), which can be considered as a good approximation of cell numbers. These RNA yields were stable across all conditions, indicating that septic injury did not result in increased release of plasmatocytes in my hands. This means that even if there is a level of heterogeneity in the larval plasmatocytes used in this thesis, this is likely to be the same across all different treatment conditions used here.

6.2 Dynamic Polycomb Chromatin

Based on the ChIP-seq data from primary plasmatocytes, in combination with the transcriptional changes observed in those plasmatocytes after immune activation, I propose to subdivide Polycomb chromatin into two distinct chromatin states: Canonical Polycomb chromatin that is characterized by high levels of H3K27me3 and found for example at epigenetically repressed developmental genes, and dynamic Polycomb chromatin that is characterized by intermediate levels of H3K27me3 and includes inactive but still inducible genes.

Since the levels of H3K27me3 signal intensity are a critical factor for subdividing Polycomb chromatin it is important to consider whether the intermediate level of H3K27me3 could represent an artifact caused by a modification different from H3K27me3 for which the antibody used in this study has a lower, but still significant affinity. In fact, LaMere, et al. [314] described that much of the genome of mouse ESCs is covered be H3K27me2, which is also placed by PRC2, and that this H3K27me2 prevents enhancer activation. This explanation is however inconsistent with the observations I made in this thesis. First, an antibody recognizing both H3K27me2 and H3K27me3 showed high enrichment in some genomic regions, but only much lower enrichment in the dynamic Polycomb chromatin state. Therefore, this chromatin state might be higher in H3K27me2 than canonical Polycomb chromatin, but it is importantly marked by the presence of H3K27me3. Further, antibodies used in this thesis are verified for their specificity for the histone modifications, and this includes the H3K27me3 antibody.[315] Here, no cross-reactivity for H3K27me2 or any other alternative histone modification is described, ruling out that the observed effect is due to a mix up of histone H3 lysine 27 modifications.

There are prior reports that use chromatin state models that include more than one H3K27me3 positive Polycomb like state. These reports, however, do not address the relevance of such a state and their assignments differ from mine in some key factors. Kharchenko, et al. [55], who build a 9 state model from a combination of *Drosophila* cell lines using HMM, find one chromatin state which has H3K27me3 levels between those states that they identify as Polycomb and active states. Throughout their work they identify these areas as background, noting that they are poor in genes or at most intronic. In contrast, my results show that dynamic Polycomb chromatin is similarly dense in genes compared to canonical Polycomb chromatin, and much denser than H3K9me3 positive heterochromatin. Kundaje, et al. [316] on the other hand uses human ChIP-seq data from the Encode consortium to build an 18 state model and they identify a distinct chromatin state, which they term weak repressed Polycomb, acknowledging that this state is likely associated with PcG proteins, and this state carries intermediate levels of H3K27me3, but is low in H3K9me3. Again, they do not address potential function that such a weak repressive Polycomb chromatin state might have. Evans, et al. [167] address chromatin states in *Caenorhabditis elegans* using data from modEncode to build a model consisting of a total of 20 chromatin states, among which they identify 3 Polycomb states. While they do not go into much detail regarding most of the states, they mention that one of the Polycomb states with intermediate H3K27me3 levels is enriched for odorant receptors, a class of genes which are only expressed in a small subset of cells. Therefore, this state might resemble the dynamic Polycomb chromatin described here targeting non-developmental genes with a high degree of temporal and special regulation.

Kharchenko, et al. [55], Kundaje, et al. [316] and Evans, et al. [317] all use Hidden Makarov Models (HMM), whereas I use a mathematically distinct binomial expectation maximization model. Therefore, it is important that both model building strategies, using gene level H3K27me3 and a combination of all histone marks, I arrive at a model that predicts the dynamic Polycomb chromatin state, which carries the intermediate levels of H3K27me3 described by others.[55,167,316] At the same time, I do observe a large number of genes in this chromatin state, and while I demonstrated that immune genes are among them, I also want to show that there is a general class of genes that are contained in dynamic Polycomb chromatin. To demonstrate this, I identified enriched Gene Ontology (GO) terms in all state models build in this thesis using a hypergeometric model. From these GO terms I selected those, which were most variant across the different chromatin states and plotted their probabilities, with circle transparency indicating p-values and circle size the effect size (Figure 43). First, active gene and enhancer chromatin states (states 4, 6, 7 and H3K27me3 low) are strongly associated with many basic cellular processes of metabolism and transcription, all listed at the top. At the bottom, there are a number of gene modules that are either related to development or generally belong to RNA, transcription or gene expression, most likely representing transcription factors. These categories are enriched in both active states and the highly repressed canonical Polycomb states (state 5 and H3K27me3 high). Interestingly, no GO terms from the top or the bottom of the list are enriched in the newly defined dynamic Polycomb chromatin state (state 3 or H3K27me3 medium), in which instead a group of GO terms in the middle of the plot is enriched. These terms appear to be GO terms which are related to more dynamically expressed genes, including signaling factors and pathways. Interestingly olfactory receptors are enriched among this dynamic Polycomb chromatin as well (for example detection of chemical stimulus involved in sensory perception of smell), which mirrors the observations for the H3K27me3 intermediate state in *C. elegans*.[317]

Given that the dynamic Polycomb chromatin contains many genes, but these genes appear to be in a functional category separate from those genes which are either constitutively active or targeted by Polycomb, it is likely that this represents a true chromatin state that cannot be simply assigned as background, as others have suggested.[55] However, since Kharchenko, et al. [55], Kundaje, et al. [316] and Evans, et al. [317] restrict themselves to building the chromatin state models, they do not further test the function of the identified states, including the H3K27me3 intermediate states. In contrast, by depleting H3K27me3 and by inducing physiological immune signaling, I showed here that the dynamic Polycomb chromatin state which I observe for immune genes and many other genes is functional in repressing genes and dynamically regulated in response to signaling.

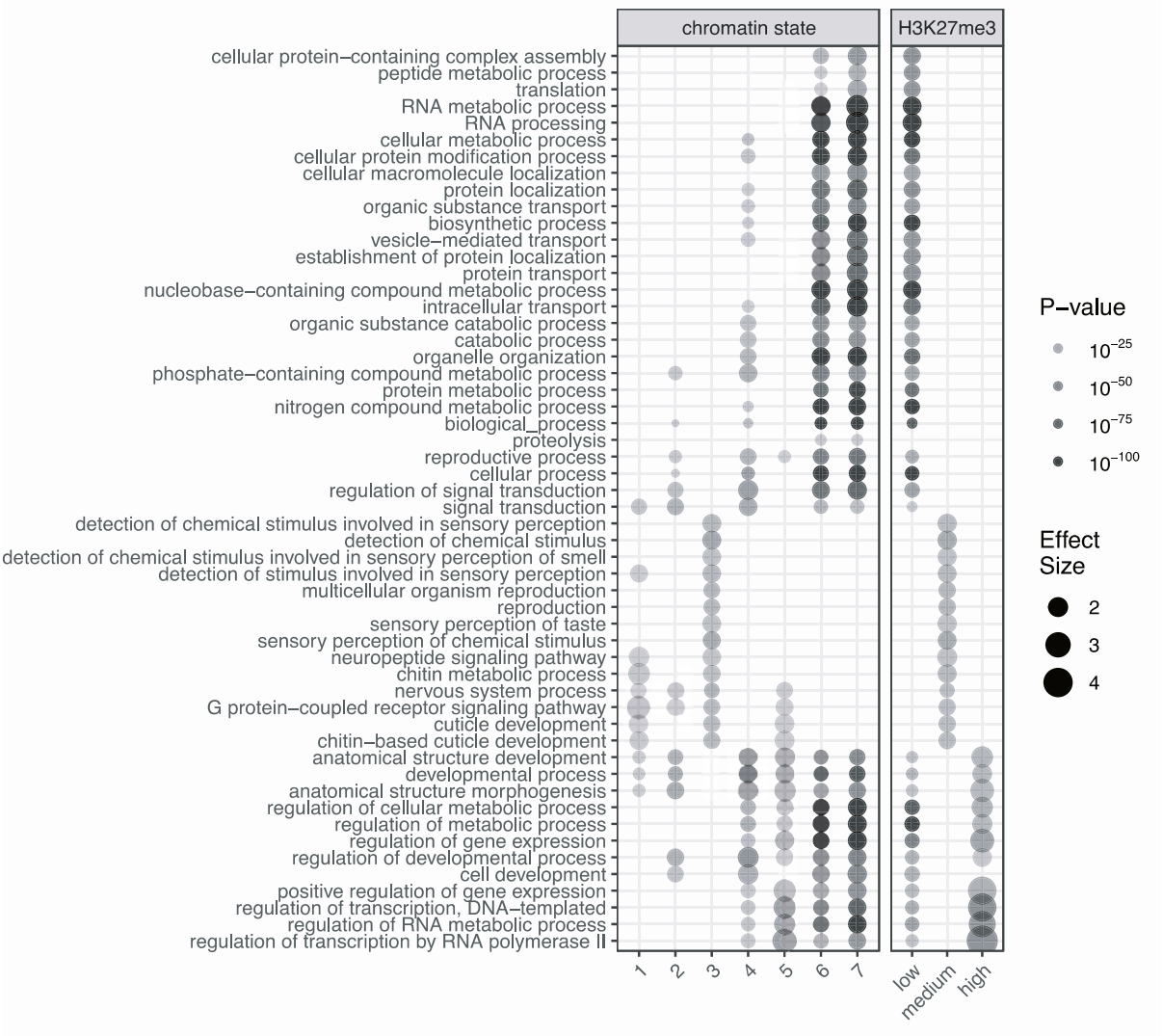

Figure 43: Genes with a high special and temporal variability are enriched on dynamic Polycomb chromatin. Genes were grouped into genes overlapping with one of the 7 chromatin states or genes with low, medium or high H3K27me3 states. For all genes GO annotation was derived from the Gene Ontology consortium[318] and enrichment was tested using a hypergeometric distribution. Among all GO terms, those with the highest variance in module associated genes were selected and plotted. Plotted were all enrichments with $p < 0{,}05$, plotting p-values as circle transparency and effect size as circle size.

6.3 Molecular basis of dynamic Polycomb chromatin

Defining two Polycomb chromatin states by the levels of H3K27me3 raises the question, how distinctly different levels of this modification could be implemented on a molecular level. Individual nucleosomes contain exactly 2 copies of $H3^{8}$, and therefore 3 different levels for individual histone modifications in one nucleosome are possible: either 0, 1 or 2 modifications per

nucleosome can be found for a specific amino acid residues. Therefore, an H3K27me3 low state could consist of nucleosomes that carry no histone H3 lysine 27 trimethylation at all, while in a H3K27me3 high state almost every nucleosome is modified twice. For dynamic Polycomb chromatin, which is characterized by intermediate levels of H3K27me3, two scenarios could then be considered. On one hand, nucleosomes could be hemi-modified homogeneously over the cell population and the two homologous chromosomes. On the other hand, the degree of modification may be variable in such chromatin regions with some nucleosomes skipped or doubly modified, but averaging out to an overall intermediate modification level.

Homogeneous hemi-modification might well explain the observed levels of H3K27me3 at gene levels, but it is difficult to imagine how such a pattern could be maintained, unless there exists a co-factor that checks modification of one H3 per nucleosome to then prevent the same modification on the second H3. While a similar but positive regulation could be implemented through the ability of Extra sex combs (ESC) to bind H3K27me3 and recruit PRC2 for further modification[114], a factor that antagonizes H3K27me3 modification analogously has not been described yet. Therefore, it is likely that the H3K27me3 intermediate chromatin state involves stochastic or sparse modification of nucleosomes with H3K27me3.

It is important to note however that ChIP methods that rely on populations of cells necessarily fail to discriminate between a scenario in which a given chromatin region is sparsely modified in all cells of the population, and a scenario in which the same region is highly modified in some cells but not at all in other cells of the population. In other words, they fail to determine at which level the stochastic element comes into play – at the chromatin level or at level of the cell population. Accordingly, in the most extreme case of population level diversity even in an ontologically homogeneous cell populations like plasmatocytes, dynamic Polycomb chromatin could not actually reflect a distinct chromatin state but in fact chromatin regions that are active in some cells and 'classically' repressed Polycomb chromatin in others. Two observations argue against this possibility. First, many genes that I assigned to dynamic Polycomb chromatin are not expressed across the population in resting plasmatocytes. This is not consistent with a significant portion of cells with non-repressed genes in these chromatin regions, given that they are up-regulated after H3K27me3 depletion and therefore actively repressed. Second, upon stimulation, H3K27me3 is actually lost from up-regulated genes. Hallmark of 'classic' Polycomb chromatin is that genes remain silenced even 'against' transcriptional input.[319] This strongly argues for dynamic Polycomb chromatin as a functionally distinct chromatin state.

A final answer, however, about how this chromatin state is implemented on the chromatin level would require high coverage single cell methods to identify histone modifications in an allelic fashion, which do not exist at the moment.

6.4 Functional dynamics of H3K27me3

Regardless of how exactly the dynamic Polycomb chromatin is established, whether it derives from a homogeneous intermediate level of H3K27me3 or a mixture of active and repressed gene states with cell-to-cell variability, it is characterized, and distinguished from constitutive Polycomb chromatin, by a dynamic removal of the H3K27me3 mark upon transcriptional activation of genes. I demonstrated that immune genes, which were targeted by H3K27me3 in unchallenged plasmatocytes, are depleted in that histone modification after septic injury concordant with their transcriptional activation. Such reduction could either represent a sporadic removal of H3K27me3 in the homogeneous H3K27me3 model, or a switching of genes in a number of cells from a H3K27me3 positive repressed state to a H3K27me3 negative active states. Either process would involve the removal of H3K27me3 from associated genes, a process which could in principle happen in one of 2 ways: Either H3K27me3 is enzymatically removed by the dedicated demethylase Utx, or the entire modified H3, perhaps even the entire nucleosome, is evicted from its position and replaced by a fresh or unmodified H3 or nucleosome. This histone exchange can happen in a number of ways, both co-transcriptionally, where RNA polymerase II recruits a number of chromatin remodelers[320], or in the absence of transcription, where for example the chaperon of the variant histone H3.3 Hira[321,322] and the SWItch/Sucrose Non-Fermentable (SWI/SNF) complex[323] have been shown to help the eviction of present histones to allow the insertion of unmodified ones.

The fact that I observe a loss in H3K27me3 is crucial, particularly in light of the cell-to-cell variability model of intermediate levels of H3K27me3. If there was a system were some plasmatocytes had their immune genes repressed by H3K27me3 while another subset had their immune genes in an active chromatin state, those cells with active immune genes could alone be responsible for the increase in immune gene transcription after septic injury, without changes in H3K27me3. Only because I can show a loss of H3K27me3 after septic injury, I can be sure that this intermediate H3K27me3 state is truly a dynamic Polycomb chromatin state.

Interestingly, I also see a gain in H3K27ac after immune stimulation. This mark is placed by the acetyltransferase *nej*, a member of the Trithorax group (TrxG), a group of proteins which has been well described to counteract the activity of proteins of the Polycomb group (PcG) in

development.[324,325] In *Drosophila*, the activity of TrxG prevents the silencing of developmental genes to which they are targeted, permitting a stable "on" gene state.[28,119] Therefore it is possible that TrxG proteins are also recruited to immune genes in plasmatocytes after septic injury to help overcoming Polycomb silencing and prevent the reestablishment of a silenced gene state. Interestingly, it has previously been demonstrated in human macrophages, that H3K27ac is differentially regulated in response to immune activation, and that this may play a role in a process that creates 'immune memory' that persists for several days.[241] Similarly, the activity of TrxG members and the placement of H3K27ac may stabilize the expression of target immune genes in plasmatocytes or increase future inducibility of those genes after a first immune activation.

6.5 Targeting of the dynamic Polycomb chromatin state

In this work I also showed that the establishment of silencing by dynamic Polycomb chromatin in principle involves the same mechanisms that are used to establish constitutive Polycomb chromatin: The PRC2 complex modifies histone H3 at lysine 27 to produce H3K27me3, which then causes gene silencing.[25] In *Drosophila* it is well described that the classic Polycomb silencing mechanism involves dedicated genetic elements, called PREs, which bind PRC2 in a sequence depended manner and are sufficient to confer silencing to adjacent genes.[122] However, so far only very few PREs have been predicted, i.e. by Ringrose, et al. [276] who predicted 167 PREs and Zeng, et al. [275] who predict several hundred PREs. This is in contrast to the several thousand of genes that I demonstrated here to be targeted by either canonical or dynamic Polycomb chromatin. This raises the question whether PREs are also involved in the silencing of dynamic Polycomb chromatin, and if they are, why they were not predicted based on a sequence motive analysis. It is possible, that PREs that target dynamically regulated genes are quite different from the well-characterized PREs at developmental genes, which would prevent sequence-based prediction. Importantly, if PREs are responsible for establishing dynamic Polycomb chromatin as well, based on the characteristics I observed here they must differ in at least one fundamental way: Classic PREs have been identified for their capacity to confer silencing to transgenic constructs in a fashion that will not be overcome by later transcriptional activation.[102] The genes I describe here are however activated by physiological transcriptional signaling. This even raises the question, whether such sequences that recruit PRC2, if they existed in dynamic Polycomb chromatin, would eventually fit the strict definition of PREs.

Many alternative models have been proposed in vertebrate systems by which Polycomb factors could be recruited to regions which are to be silenced, including long non-coding RNAs, nascent

transcripts and CpG islands.[326] One interesting model proposes that absence of gene expression itself is sufficient to cause H3K27me3 placement. Both Riising, et al. [327] and Hosogane, et al. [328] found that gene inactivity preceded the recruitment of PRC2 and the placement of H3K27me3 at relevant genes. They therefore conclude that PcG protein recruitment appears not instructive for silencing, but is a consequence of inactivity. Such observations have led to the development of the "chromatin sampling" model, where PcG proteins randomly sample chromatin regions and then modify histones in places where they are not quickly displaced by transcriptional activity.[329] It is important to point out however, that the model of gene inactivity preceding stable Polycomb silencing in no way contradicts the PRE based *Drosophila* model: It is well described that PREs are primarily memory elements that maintain silencing, but they do not carry the information required to establish the cellular pattern of silencing (which is instead formed by segment specific enhancers).[118] And while earlier reports proposed that transcriptional activity at silenced PREs would be sufficient to neutralize their activity and allow activation of adjacent genes[121], recent reports contradict this hyothesis.[319]

One factor that has frequently been discussed, particular in vertebrate development, is the idea of bivalent chromatin states, which refers to chromatin that is modified at the same time with active histone marks by TrxG proteins and with repressive marks by PcG proteins.[330] Such genes are then often described as poised for transcription, that is that they are inactive but can be activated by transcriptional signaling. In vertebrates a poised state consisting of genes which are simultaneously positive for H3K27me3 and H3K4me3 is the best described bivalent state[331], while in *Drosophila* a bivalent state with H3K27me3 and H3K4me1 has been proposed.[332] I can however not identify any H3K4me3 H3K27me3 double positive chromatin regions, ruling out such bivalent states. In my data, H3K4me1 H3K27me3 double positive regions can be seen on a small number of developmental genes, including Hox genes, where there are localized at either site of sharp RNA-PolII peaks in regions far away from transcriptional start sites, likely marking gene enhancers.[256] Interestingly, these observations are restricted to silent genes, which are, as plasmatocytes are fully differentiated cells, unlikely to ever be activated, and therefore resemble the description of Bonn, et al. [333]. At the same time those immune genes which are targeted by intermediate levels of H3K27me3 do not appear to be associated with such H3K27me3 H3K4me1 double positive enhancers, which reasons against bivalent chromatin states as a cause of transcriptional inducibility.

In contrast Sneppen and Ringrose [66] have proposed that bistability is the mode in which poised chromatin is controlled. This means that depending on external signaling either active or inactive gene states are preferred. A similar mechanism could be at work here (Figure 44), where inducible

genes are generally kept in a silenced state, which suppresses gene transcription at signaling levels close the background noise. Once a specific activation threshold is met, however, the gene turns to an active chromatin state, which is permissive of transcriptional activity, and the gene can produce transcript relative to the level of signaling input. In such a model genes in individual cells or even individual alleles could switch states with a probability dependent on the signaling levels experienced by that individual gene, therefore resulting in a non-linear switching of individual cells as for example immune signaling increases.

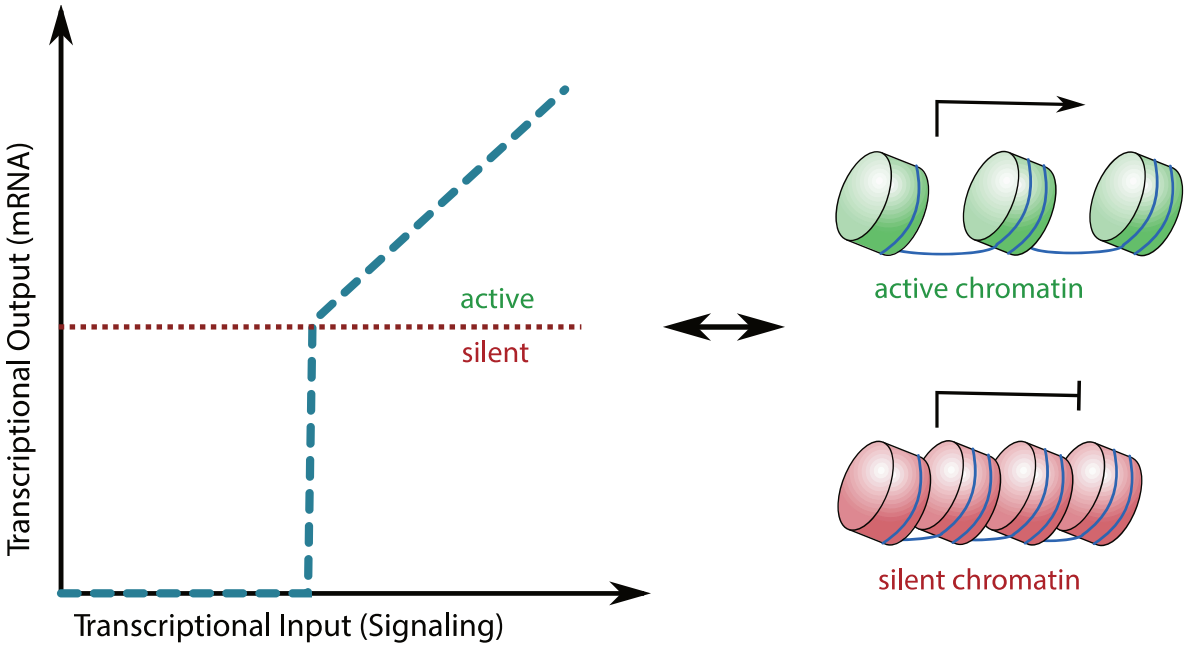

Figure 44: Model of a bistable system in dynamic Polycomb chromatin. Dynamic Polycomb chromatin (right) is repressed or silent when no or low transcriptional signaling is present, but at higher signaling levels switches to an active or permissive state. Therefore, the transcriptional output is offset by a threshold (left). In low signaling situations, no transcription can take place, and the gene does not respond to the transcription factor. Once a threshold is met, the gene turns active, and further transcriptional signaling can trigger increased transcriptional output.

6.6 Relevance of dynamic Polycomb chromatin

In the immune response two main potential uses for this non-linear behavior of dynamic Polycomb chromatin come to mind: Preventing chronic immune activation and enabling transcriptional immune memory. With regard to chronic immune activation, in the vertebrate system autoinflammation is well described.[334,335] There, it is caused by mutations or misregulation in a number of pathways and manifests itself through many symptoms including

fevers and chronic pain.[334,335] In *Drosophila* only few reports on chronic immune activation are available. Guillou, et al. [336] describe a mutant, in which chronic cytokine signaling disrupts gut homeostasis and reduces the lifespan of mutant animals, while Libert, et al. [337] demonstrate that artificial sterile activation of NFκB leads to a chronic inflammation-like state, which is marked by increased resistance against infection, but with reduced longevity in uninfected animals. Therefore, it is well possible that in the dynamic Polycomb chromatin state H3K27me3 serves to suppress low level NFκB immune activation caused by microbial patterns derived from the environment in non-infected animals. At the same time in high signaling situations, when a pathogen is encountered, the chromatin switches to an active state and transcription of effectors can proceed just as it would if the repressive state had never existed. Hence, dynamic Polycomb chromatin can help *Drosophila* to get the best of both worlds, low chronic expression of immune genes with quick and strong immune responses after and infectious challenge.

Further, immune gene targeting by dynamic Polycomb chromatin could be relevant in the transcriptional immune memory. In vertebrates, systems of innate immune memory have been described, for example in human monocytes, that allow cells to react transcriptionally differently to an immune activator when encountering it for a second time. This can lead to either an enhanced (trained) or a reduced (tolerant) reponse.[338] Histone modifications are one of the mechanism proposed by which this immune memory could be established.[241] In *Drosophila* such a mechanism that alters transcriptional immune responses to a bacterial challenge has to my knowledge not yet been described. It is nonetheless possible that dynamic Polycomb chromatin helps to establish just such a memory mechanism. Depending on how long an active chromatin state is maintained after H3K27me3 is lost, it is possible that these genes can then produce antimicrobial factors in anticipation of a future challenge or be more rapidly induced should such a second immune challenge occur.

Thus, gene regulation by dynamic Polycomb chromatin could be beneficial in many situations beyond immunity, in circumstances where incorrect or sporadic expression of genes could cause harm to the animal. Hence, I propose that dynamic Polycomb chromatin servers as a general mechanism to threshold inducible silenced genes to prevent aberrant or detrimental transcription, as I showed here for the regulation of immune genes in *Drosophila* plasmatocytes.

7 MATERIAL AND METHODS

7.1 Lab Methods

7.1.1 BACTERIAL CULTURES

Escherichia coli strain DH5α (ThermoFisher) and *Staphylococcus aureus* strain 8325-4[339,340] were picked from frozen glycerol stocks, were plated on LB agar plates and grown overnight at 37°C. The resulting plates were stored at 4°C for up to one week. Larger liquid cultures were generated by picking individual colonies from the LB agar plates and inoculating 100 ml LB medium in 500 ml Erlenmeyer flasks. The flasks were then incubated shaking at 180 rpm at 37°C overnight. To harvest the bacteria and generate bacterial pellets for septic injury, the OD_{600} of a 1:10 dilution of the overnight culture in LB was measured using the Ultrospec 2100pro. Bacterial cultures of *E. coli* and *S. aureus* were then pooled at a 1:1 OD_{600} ratio and centrifuged down at 4000xg for 10 minutes at room temperature. The resulting pellet was used for septic injury.

7.1.2 FLY RAISING

Flies were raised and maintained on standard cornmeal food at 25°C or room temperature. For experiments on wild type OrR, adult fly collections were left to lay eggs on apple juice agar plates with yeast paste overnight at 25°C. After 24h hatched larvae were washed of the plate and further larvae were allowed to hatch for 2h. Then, larvae were picked of the plate and transferred into 36mm food tubes with yeast paste at a density of 120 larvae per tube and raised at 25°C. The resulting larval collection tubes would reach 3rd instar wandering stage around 4 days after hatching, at which stage plasmatocytes were extracted.

7.1.3 STERILE AND SEPTIC INJURY

Sharpened tungsten needles were generated by electrolytically removing material from the tip of 0,25mm tungsten wire. A piece of tungsten wire was connected to a 6V AC power supply and dipped into a 10% solution of NaOH, in which the wire of the other electrode was dipped. Periodically, the wire was inspected for the progress of sharpening.

Sharpened tungsten needles were then used to perform septic injury. The wire was mounted in a Moria pin holder. Larvae were taken from food tubes by washing out all larvae with a 20% sucrose solution. Larvae were quickly washed in PBS, then dried on paper towels and transferred to a rubber dish. Larvae were then pricked with the tungsten needle either sterilized by 70% EtOH (sterile injury) or dipped into the bacterial pellet (septic injury). Larvae were then transferred back

into normal food tubes with 3 drops of PBS and incubated at 25°C for the indicated time. For infection time courses plasmatocytes were always extracted around the same time (15.00 to 18.00), and injuries were performed 3h, 6h or 18h prior.

7.1.4 Larval RNA isolation and qPCR

For RNA isolation from total larvae, they were collected 6h after treatment into 1,5 ml Eppendorf tubes and snap frozen in liquid nitrogen. Larvae were then lysed by adding 900 µl of TRIzol and disrupting them with a micropestle. Tubes were briefly centrifuged (16000g, room temperature, 2 minutes) to remove debris, then the supernatant was moved to a fresh prespun phase lock heavy tube. Next, 250 µl chloroform was added to samples, they were mixed thoroughly and centrifuged (12000g, room temperature, 15 minutes). The upper aqueous phase was then moved to a fresh DNA LoBind tube, making sure not to disrupt the interphase or phenol phase, and mixed with 550 µl isopropanol and 1 µl glycogen (20 mg/ml, RNAse free). Samples were mixed by inverting and incubated for 30 minutes in the freezer at -20°C. Afterwards samples were centrifuged (16000g, 4°C, 10 minutes) and the supernatant was removed carefully without disrupting the pellet. The pellet was resuspended in 100 µl ultra-pure water with 300 mM sodium acetate and 1 µl glycogen (20 mg/ml, RNAse free). Next, 300 µl EtOH added and the sample was incubated for 20 minutes at -20°C, then centrifuged (16000g, 4°C, 10 minutes) and the supernatant was discarded. The pellet was then washed 2 times by adding 1 ml of 70% EtOH (prepared with ultra-pure water), each time spinning down the pellet (16000g, 4°C, 3 minutes). Afterwards all supernatant was drained and the pellet was dried until no liquid was visible. The pellet was then resuspended in 15 µl of ultra-pure water.

This RNA was quantified by Nanodrop and reverse transcribed into cDNA using the QuantiTect kit according to the manufacturer's instructions. The resulting cDNA was diluted 1:4 in ultra-pure water and 2 µl of cDNA dilution per well was mixed with 10 µl Fast SYBR Green master mix, 2 µl of primers (see table below) and 6 µl ddH2O in a MicroAmp qPCR plate.

Target	Forward primer	Reverse primer
Rp49	GACGCTTCAAGGGACAGTATCTG	AAACGCGGTTCTGCATGAG
Dpt	GCTGCGCAATCGCTTCTACT	TGGTGGAGTGGGCTTCATG
Drs	CGTGAGAACCTTTTCCAATATGATG	TCCCAGGACCACCAGCAT

The plate was then sealed using MicroAmp transparent film and run on a QuantStudio3 using standard SYBR green fast run settings. From the resulting Ct values, relative expression values were calculated as:

$$relative\ expression = 2^{Ct_{rp49} - Ct_{AMP}}$$

7.1.5 PLASMATOCYTE ISOLATION FOR DIRECT RNA EXTRACTION

For plasmatocyte extraction and RNA isolation, treated or untreated 3rd instar wandering larvae were collected from food tubes and washed in PBS. At the same time Schneider's medium with FCS with 10 mM N-Acetyl-L-Cysteine was prepared in a well of a 24-well tissue culture plate (Sarstedt). Larvae were dried on paper towels and transferred to the tissue culture well with medium. Then they were ripped by holding the head with forceps (Dumont) and with another forceps ripping open the cuticula along the body of the larvae without disrupting the gut. In this ways 80 larvae were bled for each sample. The larval carcass was then removed and the plasmatocytes were allowed to attach for 10-15 minutes. Afterwards, the plasmatocytes were washed 4 times with PBS, then all supernatant was aspirated and they were lysed, either in 500 µl TRIzol or in 63 µl RNAdvance lysis buffer. TRIzol samples were stored at -80°C, while RNAdvance lysed samples were processed immediately.

7.1.6 CROSSES FOR MUTANT MOSAICS

Crosses for Su(z)12 and E(z) mosaics were set up using the crosses from Table 7 as:

$$hsFLP, tubGal4, uasGFPnls;; \frac{tubGal80, FRT80B}{TM6B, tb^1} \quad x \quad \frac{tubGal80, FRT2A}{TM3, Sb^1}$$

$$\Downarrow$$

$$\frac{hsFLP, tubGal4, uasGFPnls}{Y};; \frac{tubGal80, FRT2A}{TM6B, tb^1} \quad x \quad \frac{Su(z)12^4, FRT2A}{TM6B, tb^1} \ or \ \frac{E(z)^{731}, FRT2A}{TM6C, tb^1}$$

$$\Downarrow$$

$$hsFLP, tubGal4, uasGFPnls;; \frac{Su(z)12^4, FRT2A \ or \ E(z)^{731}, FRT2A}{tubGal80, FRT2A}$$

Crosses for histone mutants were set up using the crosses from Table 7 as:

$$hsFLP; \frac{+}{T(2;3)TSTL}; \frac{D^3}{T(2;3)TSTL} \quad x \quad \frac{\Delta His^C, FRT40A}{CyO, ftz}; 6xHisGU$$

$$\Downarrow$$

$$\frac{hsFLP}{Y}; \frac{\Delta His^C, FRT40A}{T(2;3)TSTL}; \frac{6xHisGU}{T(2;3)TSTL} \quad x \quad ubiGFP, FRT40A; 6xHisGU$$

$$\Downarrow$$

$$\frac{hsFLP}{Y}; \frac{\Delta His^C, FRT40A}{ubiGFP, FRT40A}; 6xHisGU$$

The crosses for the final larvae were left to lay eggs for 4h using 20 female and 10 male flies. The resulting larvae were then heat shocked twice, once in the 1st instar stadium (24h-48h after egg laying) and once in the 2nd instar stadium (48h-72h after egg laying). Then 3rd instar wandering non-tb female larvae were collected for further processing.

7.1.7 Plasmatocyte isolation and FACS sorting

For FACS sorting of plasmatocytes appropriate 3rd instar wandering larvae were collected. These were either OrR and Hml-dsRed larvae (see Table 7), or non-tb female larvae from mosaic crosses (see 7.1.6). These larvae were washed in PBS, dried on paper towels and transferred to Nunclon Sphera ultra-low attachment dishes filled with 0,5 ml Schneider's medium with FCS with 10 mM N-Acetyl-L-Cysteine. Larvae were then bled by holding the head with forceps (Dumont) and with another forceps ripping open the cuticula along the body of the larvae without disrupting the gut. Carcasses were then removed and the cell suspension was cell stained using 70 µm cell FlowMi cell strainers. DAPI was added to a final concentration of 1 µg/ml to mark dead cells. For all analysis plots shown in this thesis cells were then sorted using MACSquant Analyzer, dsRed was detected using 488nm excitation and 585/40 filters, GFP was detected using 488nm excitation and 525/50 filters and DAPI was detected using 405nm excitation and 450/50 filters. FACS data was analyzed using FlowJo10.

Sorting of cells for RNA isolation was performed in the Flow Cytometry Core Facility (DRFZ), using the BD FACSAria ROWLF, sorting for GFP using a 488nm laser for excitation and a 530/30 filter for emission and sorting for DAPI negative cells using a 405nm laser for excitation and a 450/40 filter for emission. The resulting cells were collected, spun down (2000g, 4°C, 5 minutes) and lysed either in 500 µl TRIzol or 63 µl RNAdvance lysis buffer. TRIzol samples were stored at -80°C, RNAdvance samples were isolated immediately.

7.1.8 RNA isolation by TRIzol

Frozen TRIzol samples were thawed at room temperature, adjuster to 900 µl with TRIzol and moved to fresh prespun phase lock heavy tube. 250 µl chloroform was added to each sample, they were mixed thoroughly and centrifuged (12000g, room temperature, 15 minutes). The upper aqueous phase was then moved to a fresh DNA LoBind tube, making sure not to disrupt the interphase or phenol phase, and mixed with 550 µl isopropanol and 1 µl glycogen (20 mg/ml, RNAse free). Samples were mixed by inverting and incubated for 30 minutes at in the freezer at -20°C. Afterwards samples were centrifuged (16000g, 4°C, 10 minutes) and the supernatant was removed carefully without disrupting the pellet. The pellet was resuspended in 100 µl ultra-pure

water with 300 mM sodium acetate and 1 µl glycogen (20 mg/ml, RNAse free). Next, 300 µl EtOH was added and the sample was incubated for 20 minutes at -20°C, then centrifuged (16000g, 4°C, 10 minutes) and the supernatant was discarded. The pellet was washed 2 times by adding 1 ml of 70% EtOH (prepared with ultra-pure water), each time spinning down the pellet (16000g, 4°C, 3 minutes). Afterwards all supernatant was drained and the pellet was dried until no liquid was visible. The pellet was then resuspended in 15 µl of ultra-pure water and stored at -80°C.

7.1.9 RNA ISOLATION BY RNADVANCE

RNA was isolated from freshly lysed plasmatocytes samples using the RNAdvance cell v2 kit following manufacturer's instructions using a DynaMag-2 magnetic rack. DNase digestion was performed as suggested in the manual using Ambion DNase. Resulting RNA samples were eluted in 20 µl ultra-pure water and stored at -80°C.

7.1.10 RNA QUALITY CONTROL

RNA concentration of RNA samples for library preparation was measured using the Qubit RNA high sensitivity kit and the Qubit 2.0 fluorometer using 2 µl of RNA sample according to manufacturer instructions where applicable. 1 µl of RNA was then diluted to 1 ng/µl and run on an Agilent bioanalyzer RNA Pico by the Max Planck Institute for Infection Biology DNA/RNA analysis core facility. If the resulting samples had the expected RNA concentrations and showed the characteristic ribosomal RNA peaks on the bioanalyzer without signs of degradation, they were used for subsequent library preparation.

7.1.11 RNA LIBRARY PREPARATION

Libraries for FACSed plasmatocytes from histone mutants and matching controls were generated by the Max Planck Genome Centre (Cologne). All other libraries were prepared by me and were made using the NEBNext Ultra II Directional RNA Library Prep with Sample Purification Beads, the NEBNext Poly(A) mRNA Magnetic Isolation Module and the NEBNext Multiplex Oligos for Illumina sets according to manufacturer's instruction. Adaptor concentrations were adjusted according to manual instructs. The number of PCR cycles was adjusted depending on starting RNA amount (see Table 8). The final libraries were eluted into 12 µl ultra-pure water and stored at -20°C.

7.1.12 PLASMATOCYTE ISOLATION AND CROSSLINKING FOR CHIP

For plasmatocyte extraction and chromatin fixation, untreated or 6h septic injury 3rd instar wandering larvae were collected from food tubes and washed in PBS. At the same time Schneider's medium FCS with 10 mM N-Acetyl-L-Cysteine was prepared in a well of a 6-well tissue culture plate (Sarstedt). Larvae were dried on paper towels and transferred to the tissue culture well with medium. Then they were ripped by holding the head with forceps (Dumont) and with another forceps ripping open the cuticula along the body of the larvae without disrupting the gut. 100 larvae were bled this way for each sample. The larval carcass was then removed and the plasmatocytes were allowed to attach for 10-15 minutes. Afterwards, the plasmatocytes were washed 4 times with PBS, then all supernatant was aspirated and plasmatocytes were fixed for 10 minutes at room temperature with 1,8% Formaldehyde in PBS. Afterwards, fixation was quenched for 3 minutes using 100 µl 2M glycine. Then cells were scraped of the dish using a cell scraper and transferred to 1,5 ml tube on ice. After a 5-minute incubation on ice, plasmatocytes were centrifuged (3000g, 4°C, 5 minutes), then washed 2x with PBS-Tx (PBS + 0,1% Triton X-100 w/v), each time spinning the cells down (3000g, 4°C, 5 minutes). After the final wash, as much supernatant as possible was removed from the pellet, without losing it, then the tube was snap frozen in liquid nitrogen and stored at -80°C until the ChIP.

7.1.13 ChIP protocol

For ChIP a number of tubes corresponding to the appropriate number of fixed larval plasmatocytes (see Table 9) was thawed on ice. All tubes were then resuspended in ice cold PBS-Tx (PBS + 0,1% Triton X-100 w/v) and pooled. If several different ChIPs were performed in one run, they derive from a common cell pool. If replicates of a ChIP were performed in one run, these samples derive from separate cell pools. Pooled fixed plasmatocytes were centrifuged down (4500g, 4°C, 10 minutes) and all supernatant was drained. Cells were then resuspended in RIPA buffer supplemented with protease inhibitor cocktail, PhosSTOP and 20 mM sodium butyrate to a final concentration of 3 larval equivalents of plasmatocytes per µl. Plasmatocytes were then dissociated using a syringe with a G26 needle by passing the cell suspension through it 10 times and afterwards incubated for 30 minutes on ice. The chromatin suspension was next split into prechilled Bioruptor microtubes at 100 µl per tube and then sonicated in a Bioruptor Pico using 30s on 30s off cycles at 4°C. Sonication was performed first for 3 cycles and then 2 times for 4 cycles with in-between vortexing of tubes. Sonicated chromatin samples were then spun down to remove unsheared debris (16000g, 4°C, 10 minutes) and the supernatant was pooled into a fresh DNA LoBind tube. Samples were then split into chilled tubes according to the required plasmatocyte counts (see Table 9) holding a small amount back as input. Volumes were adjusted

to 100 µl using RIPA and antibodies were added at appropriate concentrations (see Table 1). The chromatin antibody mixture was then incubated on a rotating wheel at 4°C overnight. The next day a 1:1 mixture of protein A and protein G Dynabeads was washed 3 times using a DynaMag-2. Then the beads were resuspended in a small volume and beads were added to the overnight samples to make a final amount corresponding to 5 µl of each protein A and G beads. Bead-antibody-chromatin samples were incubated on a rotating wheel at 4°C for 4 hours. Then room beads were pulled down using a DynaMag-2 in a cold, resuspended in fresh RIPA and washed for 10 minutes on a rotating wheel. This washing was repeated 4 times with RIPA500, once with LiCl buffer and 2 times with TE buffer, for which washing was limited to 5 minutes. Finally, the beads were resuspended in 100 µl RIPA and the same was done for the untreated input sample. Then 1 µl of RNase (10 mg/ml) was added to each sample and they were incubated for 1h at 37°C. Then 4 µl 10% SDS and 3 µl Proteinase K (20 mg/ml) was added to each sample and they were incubated overnight at 65°C shaking at 1200 rpm. The next day the magnetic beads were removed using the DynaMag-2, and DNA was isolated from the supernatant using the Zymo ChIP DNA clean and concentrator kit according to the manual. The DNA was finally eluted in 30 µl ultra-pure water and stored at -80°C.

7.1.14 ChIP quality control

ChIP samples were checked for concentration of DNA using 5 µl of ChIP sample in the Qubit DNA high sensitivity kit using the Qubit 2.0 fluorometer. Differential enrichment of genomic regions by ChIP was then confirmed by qPCR. 2 µl of each ChIP was diluted across reactions for 8 regions with 2 replicates each and qPCR was performed using the Fast SYBR Green master mix with primers targeted against well characterized genomic regions (see Table 6) according to the manufacturer's instructions. All samples passed these quality control tests.

7.1.15 ChIP library preparation

ChIP-seq libraries were prepared from the final ChIP samples using the NEBNext Ultra II DNA kit and the NEBNext Multiplex Oligos for Illumina sets according to the manual with one minor adjustment: Size selection of the library was performed after PCR amplification. For this, after PCR amplification 27,5 µl of AMpure XP beads was added to each sample and incubated for 5 minutes. Then, the tubes were placed on a DynaMag-2 and the supernatant was moved to fresh tubes, while the beads were discarded. To the supernatant 22,5 µl AMpure XP beads was added, they were incubated for 5 minutes at room temperature, then pulled down using the DynaMag-2 and the supernatant was discarded.

The beads were washed 2 times on the magnetic rack with 80% EtOH and then resuspended in 17 µl ultra-pure water. After incubating for 2 minutes at room temperature, the magnetic beads were pulled down and the supernatant containing the final libraries was moved to fresh tubes. The libraries were then stored at -20°C.

7.1.16 Library quantification and pooling

Both RNA-seq and ChIP-seq libraries were processed the same way for quality control and sequencing. First DNA concentration was determined in all libraries using the Qubit DNA high sensitivity kit. 1 µl of the library was then diluted down to 1 ng/µl using EB buffer and run on an Agilent bioanalyzer high sensitivity DNA ChIP by the Max Planck Institute for Infection Biology DNA/RNA analysis core facility. If the library was sufficiently concentrated and no adaptors were visible in the bioanalyzer tracks library concentrations were then determined using the KAPA library quantification kit. Therefore, all libraries were diluted 1:300000 and run in 2 wells of a MicroAmp plate in the QuantStudio3 according to the KAPA kit instructions. From the Ct values relative molar abundances of each library was calculated as 2^{-Ct} and from this pooling ratio volumes were determined, that would yield to desired read count for each library. Where necessary libraries were diluted in EB buffer.

7.1.17 Library sequencing

All libraries were sequenced either at Max Planck Genome Centre (Cologne) or at the Max Planck Institute for Molecular Genetics (Berlin). Which library was sequenced where and further run conditions are listed in Table 8 and Table 9.

7.1.18 Plasmatocyte isolation and immunofluorescence

For immunofluorescence of Su(z)12 and E(z) MARCM mosaics, heat shocked female non-tb 3rd instar wandering larvae of the cross described in 7.1.6 were collected from food tubes and washed in PBS. At the same time Schneider's medium FCS with 10 mM N-Acetyl-L-Cysteine was prepared in 8-well ibidi microscopy slides. Larvae were dried on paper towels and transferred to the ibidi dish well with medium. Then they were ripped by holding the head with forceps (Dumont) and with another forceps ripping open the cuticula along the body of the larvae without disrupting the gut. In this ways 20 larvae were bled for each well. The larval carcass was then removed and the plasmatocytes were allowed to attach for 10-15 minutes. Afterwards, the plasmatocytes were washed 4 times with PBS, then all supernatant was aspirated and plasmatocytes were fixed with 4% PFA in PBS for 5 minutes at room temperature. Plasmatocytes were washed 3 times with PBS-Tx (PBS + 0,1% Triton X-100 w/v), quenched for 5 minutes with PBS-Tx with 20 mM glycine and then permeabilized for 15 minutes with PBS with 0,4% Triton X-100. Next the samples were blocked in PBS-Tx with 10% goat serum for 20 minutes. Then H3K27me3 or H3K27ac antibodies were added at a 1:500 dilution (see Table 1) in PBS-Tx with 10

% goat serum. The primary antibody staining was incubated over night at 4°C and then washed 3 times for 5 minutes with PBS-Tx. In parallel, the secondary antibody (1:400, see Table 1) solution was prepared in PBS-Tx with 10% goat serum and quickly spun down to remove antibody aggregates. Plasmatocytes samples were stain with the secondary antibody solution for 1h at room temperature. Afterwards, the plasmatocytes were washed 3 times for 5 minutes with PBS-Tx, then washed for 5 minutes with PBS-Tx with 1 µg/ml DAPI, washed once with ultra-pure water and then mounted in ProLong Diamond. The plasmatocyte samples were imaged on a Leica SP8, using 405nm, 488nm and 563nm lasers at 20x magnification for 8-10 positions per well. This was repeated for 3 independent experiments. The resulting data was processed in ImageJ. First cell nuclei were automatically identified by the DAPI staining: A 0.4-micron gaussian blur was applied, then the image was thresholded using the default auto threshold method. Cells were identified by the Analyze particles algorithm using a circularity of 0.5-1 and a size of 5-40 square microns. For these regions of interest, the signal was determined for each individual picture in all 3 channels, GFP, DAPI and antibody staining. The data was then transferred to R, where a manual threshold for the GFP staining was determined and the antibody staining intensity was normalized to the DAPI intensity. This cell level normalized intensity was then used as a metric for histone modification abundance in GFP positive (mutant) and GFP negative (wild type) plasmatocytes.

7.2 NGS Bioinformatics

7.2.1 RNA DATA MAPPING AND QUALITY CONTROL

All sequencing data was transferred from the Max Planck Genome Centre (Cologne) or the Max Planck Institute for Molecular Genetics (Berlin) as fastq files that were pre-split according to library indexes with a 0 mismatch de-multiplexing. The completeness and intactness of transferred data was checked by MD5 sums. The quality of the sequencing run was determined using FastQC.

The reference genome fasta sequence file of the Berkeley Drosophila Genome Project assembly dm6 and the related gtf genome annotation file for dm6 of ensembl release 91 (dm6.91) were downloaded from ensembl (www.ensmbl.org).[341] A reference genome index was generated using dm6.91 using STAR-2.7.0e[342] calling

```
STAR --runMode genomeGenerate --runThreadN 10 --genomeDir TargetDir
--genomeFastaFiles SourceDir/Drosophila_melanogaster.BDGP6.dna.toplevel.fa
--sjdbGTFfile SourceDir/Drosophila_melanogaster.BDGP6.91.gtf
```

The resulting index was used to map the fastq files to the genome using

```
STAR --genomeDir GenomeDir –readFilesIn ForwardMate.fastq.gz ReverseMate.fastq.gz
--readFilesCommand gunzip -c --outSAMtype BAM SortedByCoordinate
--alignIntronMin 12 --outFileNamePrefix OutputDir --runThreadN 6
```

Next, an index file was generated for the resulting bam file using samtools[343]

```
samtools index TargetBam
```

And duplicate reads were removed using Picard tools

```
java -Xmx32G -jar picard-2.8.2.jar MarkDuplicates INPUT=StarOutput.bam
OUTPUT=NoDuplicateFile.bam METRICS_FILE=SampleID_MarkDuplicates_metrics.txt
OPTICAL_DUPLICATE_PIXEL_DISTANCE=2500 CREATE_INDEX=true TMP_DIR=/tmp
REMOVE_DUPLICATES=true ASSUME_SORTED=true
```

BigWig signal track files were generated from the duplicate free bam file using IGVtools[344] and UCSCtools[345]

```
igvtools count --minMapQuality 30 NoDuplicateFile.bam stdout dm6 | wigToBigWig -clip
dm6.chrom.sizes NoDuplicateFile.bw
```

These BigWig files were used to visualize genome tracks in IGV[344]

Quality control of RNA-seq mapping was further performed using RSeQC[346] applying the functions bam_stat, clipping_profile, inner_distance, infer_experiment, insertion_profile, junction_annotation, read_distribution, read_dupplication, RPKM_saturation and geneBody_coverage according to the RSeQC manual.

All quality control files for FastQC, STAR mapping, Picard tools and RSeQC were aggregated and visualized using MultiQC[347] and all data was checked to make sure the library and sequencing was of good quality.

Once a data set passed quality control, the gene level read counts were determined from duplicate free bam files using the subread package[348]

```
featureCounts -p -s 2 -t exon -g gene_id -a Drosophila_melanogaster.BDGP6.91.gtf -o RunID.count
NoDup1.bam NoDup2.bam NoDup3.bam …
```

These read counts were then used for further analysis.

7.2.2 RNA SEQ GENERAL DIFFERENTIAL EXPRESSION ANALYSIS

The gene level read counts were then loaded into R. For PCA analysis the matrix of gene level read counts was transformed using the DESeq2[349] rlog function, from which the 1000 most variant genes were selected, PCA analysis was performed using the stats package prcomp function and visualized using the plot3D package.

For differential expression analysis, gene level read counts were processed using the edgeR package with the quasi-likelihood general linear model approach[350,351] according to the manual:

```
y <- DGEList(CountMatrix, group = group, genes =geneID)
y <- y[rowSums(cpm(y)>1) >= 4, , keep.lib.sizes=FALSE]
y <- calcNormFactors(y)
y <- estimateDisp(y, design = designMatrix, robust = T)


fit <- glmQLFit(y, designMatrix, robust=TRUE)
results <- glmQLFTest(fit, contrast = ContrastOfInterest)
DifferentialTest <- decideTests.DGELRT(results, lfc = 1)
```

Here, the *results* object contains gene level $\log_2$ fold changes and p-values, while the *DifferentialTest* object determines differential expression with false discovery correction. The resulting data was plotted using ggplot2. For heatmaps DESeq2[349] rlog transformed read counts were transformed to z-scores using the R base package scale function and plotted with ggplot2. Hierarchical clustering was done using the stats hclust function on euclidian distances of gene vectors.

For GO term analysis the limma[352] function goana was used after converting the Flybase gene IDs to EntrezIDs using the ensembl 91 release and biomaRt[353]. For custom gene sets, the limma[352] function fry was used and results were visualized using ggplot 2, or an implementation of the GSEA algorithm (see source code Code Snippet 2) was used. For the GSEA algorithm genes were ranked by their log fold change and for each rank the enrichment score for a gene at rank i g_i was calculated as

$$ES(g_i) = \sum_{\substack{g_j \in S \\ j \le i}} \frac{|log_2 FC_j|}{N_R} - \sum_{\substack{g_j \in S \\ j \le i}} \frac{1}{(N - N_H)}, \quad where\ N_R = \sum_{g_j \in S} |log_2 FC_j|$$

where S are the genes in the set and N_H is the number of genes in the set. P values were then calculated by randomly assigning N_H genes to S, calculating the maximum enrichment score and determining the number of permutations with greater enrichment scores than the real set. Venn diagrams were visualized using the VennDiagram venn.diagram function.

For GO term analysis using hypergeometric testing, GO term annotations and GO term relations were downloaded from the Gene Ontology consortium.[318] Gene annotations were extended to

ancestor GO terms for each annotation, and genes were grouped as described. Hypergeometric testing was then performed using Code Snippet 4. From the resulting list elements with highest row variance of annotated elements in the target set was selected and plotted.

7.2.3 ChIP DATA MAPPING AND QUALITY CONTROL

As for RNA-seq, all sequencing data was transferred from the Max Planck Genome Centre (Cologne) or at the Max Planck Institute for Molecular Genetics (Berlin) as fastq files that were pre-split according to library indexes with a 0 mismatch de-multiplexing. The completeness and intactness of transferred data was checked by MD5 sums. The quality of the sequencing run was determined using FastQC.

The reference genome fasta sequence file of the Berkeley Drosophila Genome Project assembly dm6 and the related gtf genome annotation file for dm6 of ensembl release 91 (dm6.91) were downloaded from ensembl (www.ensmbl.org).[341] A reference genome index was generated using dm6.91 using STAR-2.7.0e[342] calling

```
STAR --runMode genomeGenerate --runThreadN 10 --genomeDir TargetDir
--genomeFastaFiles SourceDir/Drosophila_melanogaster.BDGP6.dna.toplevel.fa
--sjdbGTFfile SourceDir/Drosophila_melanogaster.BDGP6.91.gtf
```

The resulting index was used to map the fastq files to the genome using

```
STAR --genomeDir GenomeDir –readFilesIn ForwardMate.fastq.gz ReverseMate.fastq.gz
--readFilesCommand gunzip -c --outSAMtype BAM Unsorted –alignIntronMax 1
--outFileNamePrefix OutputDir --runThreadN 6
```

Next, the bam files were sorted and an index file was generated for the resulting bam file using samtools[343]

```
samtools sort TargetBam
samtools index TargetBam
```

For quality control insert size and alignment metrics were collected using Picard tools

```
java -Xmx32G -jar picard-2.8.2.jar CollectAlignmentSummaryMetrics I=StarOutput.bam
O=AlignmentMetrics.txt R=Drosophila_melanogaster.BDGP6.dna.toplevel.fa &
java -Xmx32G -jar picard-2.8.2.jar CollectInsertSizeMetrics I=StarOutput.bam O=InsertSizeMetrics.txt
P=InsertSizeHistogram.pdf
```

Duplicate reads were removed using Picard tools

```
java -Xmx32G -jar picard-2.8.2.jar MarkDuplicates INPUT=StarOutput.bam
OUTPUT=NoDuplicateFile.bam METRICS_FILE=SampleID_MarkDuplicates_metrics.txt
OPTICAL_DUPLICATE_PIXEL_DISTANCE=2500 CREATE_INDEX=true TMP_DIR=/tmp
REMOVE_DUPLICATES=true ASSUME_SORTED=true
```

All quality control files for FastQC, STAR mapping and Picard tools were aggregated and visualized using MultiQC[347] and all data was checked to make sure the library and sequencing was of good quality.

BigWig signal track files were generated from the duplicate free bam file using IGVtools[344] and UCSCtools[345]

```
igvtools count --minMapQuality 30 NoDuplicateFile.bam stdout dm6 | wigToBigWig -clip
dm6.chrom.sizes NoDuplicateFile.bw
```

These BigWig files were used to visualize genome tracks in IGV[344]

For ChIP heatmaps, PCAs and gene signal profiles, ChIP bam files were processed using a local version of Deeptools[255] according to the manual, data was transferred to R and visualized using ggplot2.

7.2.4 Expectation maximization analysis

For applying the binomial expectation maximization algorithm to gene level ChIP signals, gene level read counts from mapped duplicate free ChIP-seq bam files were counted using the subread package[348]

```
featureCounts -BCpO -Q 30 -T 4 -t exon -g gene_id -a Drosophila_melanogaster.BDGP6.91.gtf -o
RunID.count NoDup1.bam NoDup2.bam NoDup3.bam …
```

To extract bin level read counts from ChIP-seq bam files the bamsignals bamProfile function with a bin size of 200 was done in R

```
parallel::mcmapply(bamsignals::bamProfile, bampath=ChIPFilesWithPath, MoreArgs =
                                list(gr=gr,binsize = 200, paired.end="midpoint", tlenFilter = c(70, 500),
                                filteredFlag = 1024, verbose = F), mc.cores = 5, SIMPLIFY = F)
```

The resulting counts were processed into 2 data matrixes, 1 for ChIP signal and 1 for input signal, in which each row represents a gene or bin and each column represents a library or sample. These matrixes were then used to call the BinomEMwrapperParallel function (see source code in Code Snippet 1). This function performs binomial EM analysis as described in chapter 5.2.2. From the resulting fit object, group assignments were used for downstream analysis.

For the PCA analysis, from the same data matrixes that were used to call the EM algorithm bin level enrichments were calculated using sequencing counts normalized to the total number of reads as

$$Enrichment(bin_i) = \frac{s_i}{s_i + r_i} = \frac{\#\ reads\ in\ ChIP\ at\ bin_i}{\#\ reads\ in\ ChIP\ and\ input\ at\ bin_i}$$

and the transposed data matrix was subjected to PCA analysis using the stats prcomp package. From this a random selection of 3000 data points was plotted using plot3D, coloured according to the EM cluster assignment.

For assignment of gene overlap with clusters the GenomicRanges[354] findOverlap function was used with a minimum overlap of 1. Enrichment analysis was performed as described in 7.2.2.

7.2.5 Differential ChIP-seq

For differential ChIP-seq data was mapped as described in the ChIP-seq chapter (7.2.3). For differential analysis with MACS2[260] first extension sizes were called using

```
macs2 predictd -i ChIPrep1.bam ChIP2.bam
```

And the resulting extension sizes were used to call peaks using the broad peak algorithm

```
macs2 callpeak --broad -g dm -B -t ChIPrep1.bam ChIP2.bam -c Input1.bam Input2.bam -n infected --nomodel --extsize extsize
```

Then, differential regions were called using the bdgdiff function

```
macs2 bdgdiff --t1 cond1_treat_pileup.bdg --c1 cond1_control_lambda.bdg --t2 cond2_treat_pileup.bdg --c2 cond2_control_lambda.bdg --d1 cond1ControlTags --d2 cond2ControlTags -g 60 -l 120 --o-prefix diff_c1_vs_c2
```

The resulting broadPeak and diff_c1_vs_c2.bed flies were loaded into R and overlap between differential peak regions and all peaks as well as between peak regions and genes was determined using the GenomicRanges[354] findOverlap function with a minimum overlap of 1.

For analysis of differential region by the EM algorithm, data matrixes for gene level data (H3K27me3) and 500 bp downstream of TSS (H3K27ac) are generated as described in 7.2.4, but were then used in the EM algorithm such that the data matrix of ChIP-seq from one of the plasmatocyte treatment conditions replaced the input matrix. The resulting gene overlap was than determined as in 7.2.4.

Enrichment analysis was performed as described in 7.2.2, and heatmap and fingerprint plots were generated using Deeptools[255].

Quantile normalization was performed using the function from Code Snippet 1. Resulting genome profiles and heatmaps were generated as described in 7.2.3 and 7.2.2. Bootstrapped confidence intervals were generated using R hmisc smean.cl.boot function and plotted using ggplot2. Linear models of H3K27me3 and gene expression correlation were generated using the stats lm and cor.test functions and visualized using ggplot2.

7.3 Material Tables

7.3.1 Antibodies

Antigen	Species	use	dilution/ amount	Cat. No	Lot	Reference
H3K27me3	rabbit polyclonal	ChIP	0,25 µg	C15410195	A0821D	Diagenode SA, Belgium
CTD4H8 RNA polymerase II	mouse monoclonal	ChIP	1 µg	05-623	2885186	Merck Millipore, USA
H3K9ac	rabbit polyclonal	ChIP	0,3 µg	C15410004	A1435-0012D	Diagenode SA, Belgium
H3K9me3	rabbit polyclonal	ChIP	0,2 µg	C15410193	A1671-001P	Diagenode SA, Belgium
H3K4me1	rabbit polyclonal	ChIP	0,15 µg	C15410194	A1862D	Diagenode SA, Belgium
H3K27ac	rabbit polyclonal	ChIP	0,15 µg	C15410196	A1723-0041D	Diagenode SA, Belgium
H3K36me3	rabbit polyclonal	ChIP	0,15 µg	C15410192	A1845P	Diagenode SA, Belgium
H4K20me1	mouse monoclonal	ChIP	0,4 µg	C15200147	003	Diagenode SA, Belgium
H3K4me3	rabbit polyclonal	ChIP	0,1 µg	ab8580	GR240214-2	Abcam, UK
H3K27me2/me3	mouse monoclonal	ChIP	0,15 µg	39535	33311011	Active Motif, Belgium
H3K27me3	rabbit polyclonal	IF	1:500	39155	25812014	Active Motif, Belgium
H3K27ac	rabbit polyclonal	IF	1:500	39133	25812006	Active Motif, Belgium
anti-rabbit, Alexa 568	goat polyclonal	IF	1:400	A11011	1842719	Thermo Fisher Scientific, USA

Table 1: List of antibodies used in this thesis

7.3.2 Chemicals

Reagent	Ref. #	Manufacturer
1M Tris pH8,0	AM9855G	Thermo Fisher Scientific, USA
3M Sodium acetate	R1181	Thermo Fisher Scientific, USA
Agencourt AMPure XP	A63880	Beckman Coulter Inc., USA
Ambion DNase (RNAse free)	AM2222	Thermo Fisher Scientific, USA
Chloroform	C2432-1L	Honeywell Research Chemicals, Germany
cOmplete, Mini, EDTA-free Protease Inhibitor Cocktail	4693159001	Roche, Switzerland
Dynabeads Protein A	10001D	Thermo Fisher Scientific, USA
Dynabeads Protein G	10003D	Thermo Fisher Scientific, USA
Ethanol	1.00983.251	Merck Millipore, USA
Fast SYBR Green Master Mix	4385612	Thermo Fisher Scientific, USA
Fetal Bovine Serum	S0115	Merck Millipore, USA
Formaldehyde solution, 37%	F8775	Sigma-Aldrich, USA

Glycine	3790.2	Carl Roth, Germany
Glycogen, RNA grade, 20mg/ml	R0551	Thermo Fisher Scientific, USA
Isopropanol	P/7490/21	Fisher Scientific, USA
LB-Agar	X969.2	Carl Roth, Germany
LB-Medium	X968.2	Carl Roth, Germany
N-Acetyl-L-cysteine	A9165	Sigma-Aldrich, USA
Normal Goat Serum	S26-100ML	Merck Millipore, USA
Paraformaldehyde (PFA)	P6148	Sigma-Aldrich, USA
PBS Dulbecco w/o Ca	L182-50	Merck Millipore, USA
Pen Strep	15140-122	Thermo Fisher Scientific, USA
PhosSTOP	4906845001	Roche, Switzerland
ProLong Diamond	P36951	Thermo Fisher Scientific, USA
Proteinase K, 20mg/ml	26160	Thermo Fisher Scientific, USA
PureBlu DAPI Nuclear Staining Dye	1351303	Bio-Rad Laboratories, USA
RNAse, DNase free	EN0531	Thermo Fisher Scientific, USA
Schneider's Insect Medium	S0146	Sigma-Aldrich, USA
SDS Solution 10% w/v	V6551	Promega, USA
Sodium butyrate	B5887	Sigma-Aldrich, USA
Sucrose	S9378	Sigma-Aldrich, USA
Triton X-100	T8787	Sigma-Aldrich, USA
TRIzol Reagent	15596-026	Thermo Fisher Scientific, USA
Tween-20	P9416	Sigma-Aldrich, USA
Ultra-Pure Water, RNAse free, DNase free	L0020	Merck Millipore, USA

Table 2: List of chemicals used in this thesis

7.3.3 CONSUMABLES

Consumable	**Ref. #**	**Manufacturer**
1.5 ml Bioruptor Microtubes	C30010016	Diagenode SA, Belgium
5Prime Phase Lock Gel Heavy 2ml	2302830	Quantabio, USA
Cell scraper	83.1830	Sarstedt, Germany
DNA LoBind Tubes	30108051	Eppendorf AG, Germany
Eppendorf Micropestle	30120973	Eppendorf AG, Germany
FlowMi cell strainers, 70µm, sterile	136800070	Sigma-Aldrich, USA
MicroAmp Fast 96-Well reaction plate	4346907	Thermo Fisher Scientific, USA
MicroAmp Optical Adhesive Film	4311971	Thermo Fisher Scientific, USA
Nunclon Sphera 35mm Dish	174943	Thermo Fisher Scientific, USA
Qubit assay tubes	Q32856	Thermo Fisher Scientific, USA
Syringe with Sub-Q Needle G26	305501	Becton Dickinson, USA
TC-plate 24 Well Standard,F	833.922	Sarstedt, Germany
Tungsten wire, 0,25mm	10408	Alfa Aesar, USA
µ-Slide 8 Well	80826	ibidi GmbH, Germany

Table 3: List of consumables used in this thesis

7.3.4 KITS

Kit	Ref. #	Manufacturer
Agencourt RNAdvance Cell v2 Kit	A47942	Beckman Coulter Inc., USA
Agilent High Sensitivity DNA Kit	5067-4626	Agilent Technologies, USA
Agilent RNA 6000 Pico Kit	5067-1513	Agilent Technologies, USA
ChIP DNA Clean & Concentrator	D5205	Zymo Research, USA
Kapa Library Quant Kit (Illumina)	7960140001	Roche, Switzerland
NEBNext Multiplex Oligos for Illumina (Index Primers Set 1)	E7335	New England Biolabs, USA
NEBNext Multiplex Oligos for Illumina (Index Primers Set 2)	E7500	New England Biolabs, USA
NEBNext Multiplex Oligos for Illumina (Index Primers Set 3)	E7710	New England Biolabs, USA
NEBNext Poly(A) mRNA Magnetic Isolation Module	E7490	New England Biolabs, USA
NEBNext Ultra II Directional RNA Library Prep with Sample Purification Beads	E7765	New England Biolabs, USA
NEBNext Ultra II DNA Library Prep with Sample Purification Beads	E7103	New England Biolabs, USA
QuantiTect Reverse Transcription Kit	205311	Qiagen, Germany
Qubit dsDNA HS Assay Kit	Q32854	Thermo Fisher Scientific, USA
Qubit RNA HS Assay Kit	Q32852	Thermo Fisher Scientific, USA

Table 4: List of kits used in this thesis

7.3.5 INSTRUMENTS

Instrument	Manufacturer
DynaMag-2	Thermo Fisher Scientific, USA
NEBNext Magnetic Separation Rack	New England Biolabs, USA
Dumont #5 Forceps, Biology	Dumont Tools, Switzerland
Nanodrop 2000 Spectrophotometer	Thermo Fisher Scientific, USA
Moria Nickel Plated Pin Holder	Fine Science Tools, Germany
Bioruptor Pico	Diagenode SA, Belgium
QuantStudio3	Thermo Fisher Scientific, USA
MACSQuant Analyzer	Miltenyi Biotec, Germany
Ultrospec 2100pro	Harvard Bioscience, USA
Qubit 2.0 Fluorometer	Thermo Fisher Scientific, USA

TSC SP8 Confocal microscope	Leica, Germany

Table 5: List of instruments used in this thesis

7.3.6 qPCR primer for ChIP test

Target	Forward primer	Reverse primer	Chromatin region
abd-A	TTTCCCTGTAAGCCAGCCAG	CCAGAGCCTCGATCCAATCC	repressed
bxd	GGCCAAATTATCCGAACGCC	CTACATCCCATCCGCTTCCC	repressed
dimm	AAGGTGGGCGAAGGCTATTC	AGCCCGACAGATGAGAGAGT	repressed
Ubx	CGCAGCGATAAAACCGAAGG	TTGCCAGCTCAGCAGATGAA	repressed
Gapdh1	GCGTCGAACACAGACGAATG	TGGATCTTACCGTCCGCTTG	active
Hml	GCAAGTGTCCTTTGGCTGTG	CCGTGCTGGTTACACTCCTT	active
rpl13	TGGTCCATTCCACTTCCGTG	ATCATGGAACTGCGCCTCAA	active
rpl32	CGGCTGCCTAAGAGAGTGAG	CCTGCACTGTTCGAGTGGAT	active

Table 6: qPCR primer used for testing ChIP enrichment

7.3.7 Media and buffer

Fly food (25l):

Agar-Agar	125 g
Yeast	360 g
Soy flour	200 g
Yellow cornmeal	1600 g
Light malt extract	1600 g
Golden syrup	440 g
Propionic acid	150 ml
Nipagin	20 g

Apple juice agar (500 ml):

Agar-Agar	10 g
Sucrose	12,5 g
Apple juice	125 ml
Nipagin	1 g

Schneider's medium with FCS:

Schneider's medium	500 ml
Fetal Bovine Serum	50 ml
Pen/Strep	5 ml

RIPA buffer:

10 mM Tris-HCl (pH 8.0)
140 mM NaCl
1 mM EDTA
1% (v/v) Triton X-100
0.1% (w/v) SDS
0.1% (w/v) sodium deoxycholate

RIPA 500 buffer:

10 mM Tris-HCl (pH 8.0)
500 mM NaCl
1 mM EDTA
1% (v/v) Triton X-100
0.1% (w/v) SDS

0.1% (w/v) sodium deoxycholate

LiCl buffer:

10 mM Tris-HCl (pH 8.0)
250 mM LiCl
1 mM EDTA

0.5% (v/v) IGEPAL CA-630
0.5% (w/v) sodium deoxycholate

TE buffer:

10 mM Tris-HCl (pH 8.0)
1 mM EDTA

EB buffer:

10 mM Tris-HCl (pH 8.0)
0,05% Tween-20

7.3.8 FLY STOCKS

Genotype	Description	Reference
OrR	reference stock (wild type)	As used in modENCODE[355]
RelE20	Mutant of imd signalling	Bloomington, 55714
Rel[E20] spz[4]/TM6C, Sb[1] Tb[1]	Mutant of imd and toll signalling	Bloomington, 55718
w*; P{Hml-dsRed.Δ}/SM6a	plasmatocytes and crystal cells are marked in red	Clark, et al. [247]
P{ry[+t7.2]=hsFLP}1, y[1] w[*]; D3/T(2;3)TSTL, CyO: TM6B, Tb[1]	FLP expression after heat shock, used for histone mutants	Dr. Alf Herzig
w*; Df(2L)HisC, P{neoFRT}40A/CyO, P{ry[+t7.2]=ftz-lacC}USC1; M{3xHisGU.wt}ZH-86Fb, Pbac{3xHisGU.wt}VK00033	ΔHisC, rescued with wild type HisGU	Dr. Alf Herzig
w*; Df(2L)HisC, P{neoFRT}40A/CyO, P{ry[+t7.2]=ftz-lacC}USC1; M{3xHisGU.H3K27R}ZH-86Fb, Pbac{3xHisGU.H3K27R}VK00033	ΔHisC, rescued with HisGUH3K27R	Dr. Alf Herzig
w*; P{w[+mC]=Ubi-GFP(S65T)nls}2L P{ry[+t7.2]=neoFRT}40A; M{3xHisGU.wt}ZH-86Fb, Pbac{3xHisGU.wt}VK00033	GFP marked FRT40A with wild type HisGU, used for histone mutants	Dr. Alf Herzig
w*; P{w[+mC]=Ubi-GFP(S65T)nls}2L P{ry[+t7.2]=neoFRT}40A; M{3xHisGU.H3K27R}ZH-86Fb, Pbac{3xHisGU.H3K27R}VK00033	GFP marked FRT40A with HisGUH3K27R, used for histone mutants	Dr. Alf Herzig
P{ry[+t7.2]=hsFLP}1, P{w[+mC]=tubP-GAL4}1, P{w[+mC]=UAS-GFP.T:Myc.T:nls2}1, y[1] w[*]; RpS17[4] P{w[+t*] ry[+t*]=white-un1}70C P{w[+mC]=tubP-GAL80}LL9 P{ry[+t7.2]=neoFRT}80B/TM6B, Tb[1]	MARCM stock used for Su(z)12 and E(z) mosaics	Bloomington, 42732
y[1] w[*]; P{w[+mC]=tubP-GAL80}LL9 P{w[+mW.hs]=FRT(w[hs])}2A/TM3, Sb[1]	tubGal80 marked FRT2A, used for MARCM mosaics of Su(z)12 and E(z)	Bloomington, 5190
w*; E(z)731 P{1xFRT.G}2A/TM6C, Sb1 Tb1	E(z) mutant for mosaics	Bloomington, 24470
w*;;Su(z)12^{2},e,P{$w_{+mW.hs}$=FRT(w^{hs})}2A/TM6B, Hu, tb	Su(z)12 mutants for mosaics	Derived from Bloomington, 24469

Table 7: List of Drosophila stocks used in this thesis. Bloomington: Bloomington Drosophila Stock Center, IN, USA. Dr. Alf Herzig, Max Planck Institute for Infection Biology, Berlin, Germany.

7.3.9 RNA-SEQ LIBRARIES

Condition	Sample isolation date	RNA isolation method	Cell #	Yield (ng)	Library prep date	Library kit	PCR cycle #	Library yield (ng)	Sequencing Location	Run ID	Sample ID	Run condition	# reads (M)	% mapped	% unique
6h infection	20.09.17	RNAdvance	n.d.	177	16.10.17	NEB Next Ultra II Directional RNA	11	1200	MPGC	3381	81783	2x75	8,9	98,9	71,9
control	20.09.17	RNAdvance	n.d.	173	16.10.17	NEB Next Ultra II Directional RNA	11	1020	MPGC	3381	81784	2x75	16,5	98,9	61,5
sterile injury	20.09.17	RNAdvance	n.d.	213	16.10.17	NEB Next Ultra II Directional RNA	11	1110	MPGC	3381	81785	2x75	11,2	98,6	67,8
18h infection	21.09.17	RNAdvance	n.d.	197	16.10.17	NEB Next Ultra II Directional RNA	11	1200	MPGC	3381	81786	2x75	9,5	98,7	68,9
6h infection	21.09.17	RNAdvance	n.d.	155	16.10.17	NEB Next Ultra II Directional RNA	11	627	MPGC	3381	81787	2x75	8,5	98,7	67,9
control	21.09.17	RNAdvance	n.d.	204	16.10.17	NEB Next Ultra II Directional RNA	11	1200	MPGC	3381	81788	2x75	5,8	98,6	69,9
sterile injury	21.09.17	RNAdvance	n.d.	161	18.10.17	NEB Next Ultra II Directional RNA	11	1200	MPGC	3381	81789	2x75	8,3	98,7	70,0
18h infection	22.09.17	RNAdvance	n.d.	231	18.10.17	NEB Next Ultra II Directional RNA	11	1200	MPGC	3381	81790	2x75	8,8	98,7	70,4
18h infection	22.09.17	RNAdvance	n.d.	188	18.10.17	NEB Next Ultra II Directional RNA	11	1200	MPGC	3381	81791	2x75	7,8	98,8	67,3
3h infection	22.09.17	RNAdvance	n.d.	173	18.10.17	NEB Next Ultra II Directional RNA	11	1160	MPGC	3381	81792	2x75	6,0	98,5	69,7
control	22.09.17	RNAdvance	n.d.	124	18.10.17	NEB Next Ultra II Directional RNA	11	242	MPGC	3381	81793	2x75	7,6	97,6	61,8
18h infection	25.09.17	RNAdvance	n.d.	171	18.10.17	NEB Next Ultra II Directional RNA	11	998	MPGC	3381	81794	2x75	7,9	98,4	69,8
3h infection	25.09.17	RNAdvance	n.d.	87,1	08.11.17	NEB Next Ultra II Directional RNA	11	427,5	MPGC	3381	81795	2x75	14,2	98,5	62,8
sterile injury	25.09.17	RNAdvance	n.d.	111	08.11.17	NEB Next Ultra II Directional RNA	11	1036,5	MPGC	3381	81796	2x75	13,5	98,5	66,6

Condition	Sample isolation date	RNA isolation method	Cell #	Yield (ng)	Library prep date	Library kit	PCR cycle #	Library yield (ng)	Sequencing Location	Run ID	Sample ID	Run condition	# reads (M)	% mapped	% unique
3h infection	26.09.17	RNAdvance	n.d.	81,9	08.11.17	NEB Next Ultra II Directional RNA	11	357	MPGC	3381	81797	2x75	7,7	98,7	69,8
6h infection	26.09.17	RNAdvance	n.d.	146	08.11.17	NEB Next Ultra II Directional RNA	11	837	MPGC	3381	81798	2x75	12,9	98,8	66,3
control	26.09.17	RNAdvance	n.d.	164	08.11.17	NEB Next Ultra II Directional RNA	11	783	MPGC	3381	81799	2x75	7,1	98,7	69,8
3h infection	09.10.17	RNAdvance	n.d.	113	08.11.17	NEB Next Ultra II Directional RNA	11	631,5	MPGC	3381	81800	2x75	9,6	98,8	66,1
6h infection	09.10.17	RNAdvance	n.d.	179	08.11.17	NEB Next Ultra II Directional RNA	11	894	MPGC	3381	81801	2x75	7,2	98,8	71,5
sterile injury	09.10.17	RNAdvance	n.d.	171	08.11.17	NEB Next Ultra II Directional RNA	11	1006,5	MPGC	3381	81802	2x75	8,9	98,7	68,7
su(z)12 GFP-	24.01.18	RNAdvance	62000	n.d.	01.02.18	NEB Next Ultra II Directional RNA	13	1340	MPGC	3506	A1	2x150	8,8	94,0	80,0
E(z) GFP+	24.01.18	RNAdvance	6700	n.d.	01.02.18	NEB Next Ultra II Directional RNA	13	64,8	MPGC	3506	A2	2x150	18,2	91,5	70,7
su(z)12 GFP+	24.01.18	RNAdvance	4700	n.d.	01.02.18	NEB Next Ultra II Directional RNA	13	45,6	MPGC	3506	A3	2x150	14,1	90,2	83,2
E(z) GFP-	24.01.18	RNAdvance	28000	n.d.	01.02.18	NEB Next Ultra II Directional RNA	13	618	MPGC	3506	A4	2x150	8,5	93,9	74,3
E(z) GFP+	29.01.18	RNAdvance	4600	n.d.	01.02.18	NEB Next Ultra II Directional RNA	13	33,3	MPGC	3506	A5	2x150	15,5	90,2	77,1
E(z) GFP-	29.01.18	RNAdvance	20000	n.d.	01.02.18	NEB Next Ultra II Directional RNA	13	325	MPGC	3506	A6	2x150	10,8	93,9	68,1
su(z)12 GFP+	29.01.18	RNAdvance	3700	n.d.	01.02.18	NEB Next Ultra II Directional RNA	13	37,5	MPGC	3506	A7	2x150	15,5	90,5	81,4
su(z)12 GFP-	29.01.18	RNAdvance	37000	n.d.	01.02.18	NEB Next Ultra II Directional RNA	13	766	MPGC	3506	A8	2x150	11,6	94,0	61,9

Condition	Sample isolation date	RNA isolation method	Cell #	Yield (ng)	Library prep date	Library kit	PCR cycle #	Library yield (ng)	Sequencing Location	Run ID	Sample ID	Run condition	# reads (M)	% mapped	% unique
E(z) GFP+	31.01.18	RNAdvance	7000	n.d.	08.02.18	NEB Next Ultra II Directional RNA	13	203	MPGC	3506	A9	2x150	12,5	91,7	79,3
E(z) GFP-	31.01.18	RNAdvance	38000	n.d.	08.02.18	NEB Next Ultra II Directional RNA	13	1120	MPGC	3506	A10	2x150	12,3	93,3	55,2
su(z)12 GFP+	31.01.18	RNAdvance	6600	n.d.	08.02.18	NEB Next Ultra II Directional RNA	13	158	MPGC	3506	A11	2x150	13,9	90,3	80,9
su(z)12 GFP-	31.01.18	RNAdvance	62000	n.d.	08.02.18	NEB Next Ultra II Directional RNA	13	1690	MPGC	3506	A12	2x150	8,0	94,5	78,6
E(z) GFP+	05.02.18	RNAdvance	4000	n.d.	08.02.18	NEB Next Ultra II Directional RNA	13	206	MPGC	3506	A13	2x150	12,3	92,3	53,6
E(z) GFP-	05.02.18	RNAdvance	30000	n.d.	08.02.18	NEB Next Ultra II Directional RNA	13	1270	MPGC	3506	A14	2x150	13,0	93,9	76,2
su(z)12 GFP+	05.02.18	RNAdvance	3000	n.d.	08.02.18	NEB Next Ultra II Directional RNA	13	132	MPGC	3506	A15	2x150	15,7	91,6	35,3
su(z)12 GFP-	05.02.18	RNAdvance	40000	n.d.	08.02.18	NEB Next Ultra II Directional RNA	13	1350	MPGC	3506	A16	2x150	10,6	94,0	78,9
E(z) GFP+	06.02.18	RNAdvance	4000	n.d.	15.02.18	NEB Next Ultra II Directional RNA	13	126	MPGC	3506	A17	2x150	16,3	86,1	62,8
E(z) GFP-	06.02.18	RNAdvance	20000	n.d.	15.02.18	NEB Next Ultra II Directional RNA	13	1430	MPGC	3506	A18	2x150	12,2	93,0	62,2
su(z)12 GFP+	06.02.18	RNAdvance	3000	n.d.	15.02.18	NEB Next Ultra II Directional RNA	13	94,5	MPGC	3506	A19	2x150	17,3	90,2	78,5
su(z)12 GFP-	06.02.18	RNAdvance	37000	n.d.	15.02.18	NEB Next Ultra II Directional RNA	13	1350	MPGC	3506	A20	2x150	13,8	94,1	56,4
E(z) GFP+	07.02.18	RNAdvance	7000	n.d.	15.02.18	NEB Next Ultra II Directional RNA	13	109	MPGC	3506	A21	2x150	18,7	90,4	76,5
E(z) GFP-	07.02.18	RNAdvance	18000	n.d.	15.02.18	NEB Next Ultra II Directional RNA	13	854	MPGC	3506	A22	2x150	14,6	92,1	59,0

Condition	Sample isolation date	RNA isolation method	Cell #	Yield (ng)	Library prep date	Library kit	PCR cycle #	Library yield (ng)	Sequencing Location	Run ID	Sample ID	Run condition	# reads (M)	% mapped	% unique
su(z)12 GFP+	07.02.18	RNAdvance	2750	n.d.	15.02.18	NEB Next Ultra II Directional RNA	13	34,5	MPGC	3506	A23	2x150	17,8	87,1	76,8
su(z)12 GFP-	07.02.18	RNAdvance	27000	n.d.	15.02.18	NEB Next Ultra II Directional RNA	13	55,1	MPGC	3506	A24	2x150	9,1	92,9	75,0
H3 wt	02.11.16	TRIzol	n.d.	48,2	MPGC				MPGC	2537	A1	2x150	10,0	84,1	72,5
H3 K27R	02.11.16	TRIzol	n.d.	74,7	MPGC				MPGC	2537	A2	2x150	9,7	82,7	73,5
H3 wt	07.11.16	TRIzol	n.d.	49,5	MPGC				MPGC	2537	A3	2x150	10,2	81,2	69,5
H3 K27R	07.11.16	TRIzol	n.d.	40,8	MPGC				MPGC	2537	A4	2x150	9,1	82,7	75,1
H3 wt	08.11.16	TRIzol	n.d.	65,7	MPGC				MPGC	2537	A5	2x150	8,5	81,6	75,0
H3 wt	09.11.16	TRIzol	n.d.	53,6	MPGC				MPGC	2537	A6	2x150	9,0	80,1	71,5
H3 K27R	08.11.16	TRIzol	n.d.	68,2	MPGC				MPGC	2537	A7	2x150	10,4	80,9	71,3
H3 K27R	09.11.16	TRIzol	n.d.	124	MPGC				MPGC	2537	A8	2x150	9,1	79,9	74,0
H3 wt	14.11.16	TRIzol	n.d.	77,4	MPGC				MPGC	2537	A9	2x150	9,5	81,5	70,8
H3 K27R	14.11.16	TRIzol	n.d.	73	MPGC				MPGC	2537	A10	2x150	8,8	79,8	72,2

Table 8: List of RNA-Seq libraries generated in this thesis.

7.3.10 ChIP-seq libraries

Condition	Antibody	# Bled larvae	ChIP date (finish)	Yield (ng)	Library prep date	Library kit	PCR cycle #	Library yield (ng)	Sequencing Location	Run ID	Sample ID	Run condition	# reads(M)	% mapped	% unique
untreated	Input	105	17.08.16	36,60	22.08.16	NEB Next Ultra II DNA	12	1296	MPIMG	CH_075	538	2x50	63,77	85,7	96,0
untreated	H4K20me1	525	17.08.16	0,56	22.08.16	NEB Next Ultra II DNA	12	244,8	MPIMG	CH_075	539	2x50	49,95	66,0	79,2
untreated	H3K4me3	350	17.08.16	2,53	22.08.16	NEB Next Ultra II DNA	12	1027,2	MPIMG	CH_075	541	2x50	19,31	79,9	97,1
untreated	H3K36me3	350	17.08.16	2,37	22.08.16	NEB Next Ultra II DNA	12	732	MPIMG	CH_075	542	2x50	52,19	90,2	92,7
untreated	H3K27me2/me3	120	17.08.16	0,49	22.08.16	NEB Next Ultra II DNA	12	240	MPIMG	CH_075	543	2x50	48,95	68,7	78,1
untreated	H3K4me1	120	17.08.16	0,97	22.08.16	NEB Next Ultra II DNA	12	400,8	MPIMG	CH_075	544	2x50	39,40	86,5	90,9
untreated	Input	105	03.08.16	79,13	04.08.16	NEB Next Ultra II DNA	12	751,2	MPIMG	CH_077	556	2x50	17,66	68,9	98,7
untreated	H4K20me1	525	03.08.16	1,98	04.08.16	NEB Next Ultra II DNA	12	110,64	MPIMG	CH_077	557	2x50	7,20	63,7	96,0
untreated	H3K4me3	350	03.08.16	0,64	04.08.16	NEB Next Ultra II DNA	12	29,28	MPIMG	CH_077	559	2x50	2,26	77,3	99,5
untreated	H3K36me3	350	03.08.16	1,58	04.08.16	NEB Next Ultra II DNA	12	102	MPIMG	CH_077	560	2x50	10,61	73,4	98,0
untreated	H3K27me2/me3	120	03.08.16	3,36	04.08.16	NEB Next Ultra II DNA	12	226,8	MPIMG	CH_077	561	2x50	8,09	56,9	96,2
untreated	H3K4me1	120	03.08.16	1,39	04.08.16	NEB Next Ultra II DNA	12	71,28	MPIMG	CH_077	562	2x50	4,55	68,7	98,5
untreated	Input	90	01.03.17	34,60	20.04.17	NEB Next Ultra II DNA	13	578,4	MPIMG	CH_098	802	2x50	39,40	82,1	95,7
untreated	H3K27ac	510	01.03.17	0,85	20.04.17	NEB Next Ultra II DNA	13	169,8	MPIMG	CH_098	803	2x50	19,57	92,9	93,9
untreated	Input	90	09.03.17	52,00	20.04.17	NEB Next Ultra II DNA	13	810	MPIMG	CH_098	804	2x50	43,18	84,6	97,1
untreated	H3K27ac	510	09.03.17	0,51	20.04.17	NEB Next Ultra II DNA	13	156	MPIMG	CH_098	805	2x50	23,66	91,4	90,7
untreated	Input	30	25.08.17	40,00	28.08.17	NEB Next Ultra II DNA	12	1200	MPIMG	CH_105	897	2x50	14,81	86,4	98,7

Condition	Antibody	# Bled larvae	ChIP date (finish)	Yield (ng)	Library prep date	Library kit	PCR cycle #	Library yield (ng)	Sequencing Location	Run ID	Sample ID	Run condition	# reads(M)	% mapped	% unique
untreated	H3K9ac	210	25.08.17	0,40	28.08.17	NEB Next Ultra II DNA	12	179	MPIMG	CH_105	898	2x50	34,07	89,6	84,8
untreated	H3K9me3	210	25.08.17	1,95	28.08.17	NEB Next Ultra II DNA	12	365	MPIMG	CH_105	899	2x50	65,50	70,0	84,7
untreated	4H8CTD	450	25.08.17	0,53	28.08.17	NEB Next Ultra II DNA	12	357	MPIMG	CH_105	900	2x50	31,17	82,3	91,7
untreated	Input	30	08.09.17	20,15	13.09.17	NEB Next Ultra II DNA	12	1030	MPIMG	CH_105	901	2x50	4,57	85,6	99,1
untreated	H3K9ac	270	08.09.17	0,76	13.09.17	NEB Next Ultra II DNA	12	478	MPIMG	CH_105	902	2x50	16,37	90,3	97,4
untreated	H3K9me3	150	08.09.17	1,53	13.09.17	NEB Next Ultra II DNA	12	418	MPIMG	CH_105	903	2x50	31,44	66,5	94,4
untreated	4H8CTD	450	08.09.17	0,93	13.09.17	NEB Next Ultra II DNA	12	471	MPIMG	CH_105	904	2x50	61,41	84,1	92,2
untreated	H3K27me3	390	09.08.18	1,30	14.08.18	NEB Next Ultra II DNA	9	171	MPGC	3772	94326	2x75	76,90	84,4	73,8
untreated	Input	10	09.08.18	5,30	14.08.18	NEB Next Ultra II DNA	9	1100	MPGC	3772	94327	2x75	73,96	83,2	76,4
untreated	H3K27me3	390	09.08.18	2,20	14.08.18	NEB Next Ultra II DNA	9	206	MPGC	3772	94328	2x75	106,78	85,2	72,2
untreated	Input	10	09.08.18	4,98	14.08.18	NEB Next Ultra II DNA	9	1120	MPGC	3772	94329	2x75	66,09	84,1	78,7
6h septic injury	Input	10	30.11.18	4,11	12.12.18	NEB Next Ultra II DNA	9	1520	MPGC	4054	A1	2x150	28,20	94,0	94,0
6h septic injury	Input	10	30.11.18	6,53	12.12.18	NEB Next Ultra II DNA	9	2168	MPGC	4054	A2	2x150	24,63	93,0	93,0
6h septic injury	H3K27ac	260	30.11.18	3,42	12.12.18	NEB Next Ultra II DNA	9	1140	MPGC	4054	A3	2x150	24,72	93,5	93,5
6h septic injury	H3K27ac	260	30.11.18	4,06	12.12.18	NEB Next Ultra II DNA	9	1200	MPGC	4054	A4	2x150	26,23	93,1	93,1
6h septic injury	H3K27me3	530	30.11.18	4,70	12.12.18	NEB Next Ultra II DNA	9	610	MPGC	4054	A5	2x150	31,58	88,7	88,7
6h septic injury	H3K27me3	530	30.11.18	4,78	12.12.18	NEB Next Ultra II DNA	9	710	MPGC	4054	A6	2x150	23,04	91,1	91,1

Table 9: List of ChIP-seq libraries generated in this study

7.4 Source code

7.4.1 Expectation Maximization

```
library(Rfast)
library(foreach)
library(doParallel)
library(doSNOW)


# Algorithms for Parameter Estimation in Normal Distibutions ----------

piEst = function(E){
  require(Rfast)
  # function to calculate pi values in EM model
  # E is n by k matrix (n: number of data points, k: number of cluster groups)
  colsums(E)/dim(E)[1]
}


# Algorithms for Parameter Estimation and Expectation in Binomial Distribution ----------

BinomLnTheta = function(N, S, E){log(crossprod(S,E)) - log(crossprod(N,E))}

BinomLnA = function(N, S, lnT, empi){
  if(is.vector(N)){
    nSamples = length(N)
  }else{
    nSamples = dim(N)[1]
  }
  S %*% lnT + (N-S) %*% log(1-exp(lnT)) + matrix(rep(log(empi), nSamples), nrow = nSamples, byrow = T)
}


LnARowSums = function(lnA){
  require(Rfast)
  # lnA is n by k matrix as calculated by the lnA type functions (n: number of data points, k: number of cluster
groups)
  lnARmax <- rowMaxs(lnA, value = T)
  lnARSum <- lnARmax + log(rowsums(exp(lnA - lnARmax)))
}

# Wrappers for Binomials ----------------------------------------------------

BinomMultiDimEMcycle = function(N, S, k, maxiter = 1000, maxtol = 1E-20){
  require(Rfast)
  # X is vector of length n (n: number of data points)
  # k is number of components/models to by fitted
  # maxiter and maxtol are thresholds for exiting the optimization

  nDim = dim(N)[2]
  nSamples = dim(N)[1]

  # start bby initilazing a potential E
  lnT <- runif(k, 0, 1)
  lnT <- lnT[order(lnT)]
  lnT <- matrix(rep(lnT, nDim), nrow = nDim, byrow = T)
  lnT = log(lnT)


  empi = runif(k)
  empi = empi/sum(empi)

  # setting likelyhood of initial (random) model
  postlnL = NA
  # starting loop to find optimum
```

```
  niter = 0
  con = F

  while(!con){

    # writing over some variables for comparison
    niter = niter + 1
    priorlnL = postlnL

    # Ex step
    lnA = BinomLnA(N, S, lnT, empi)
    lnArs = LnARowSums(lnA)
    E = exp(lnA - lnArs)

    # Max step
    lnT = BinomLnTheta(N, S, E)
    empi = piEst(E)


    # checking for likelyhood convergence
    postlnL = sum(lnArs)
    if(niter > 5){
      if((postlnL - priorlnL)^2 < maxtol){
        converged = T
        con = T
        return(list(Expectaion = E, BinomPropabilites = exp(lnT), GroupSizePi = empi, ModelLnLikelyhood =
postlnL, converged = converged))
      }
    }
    if(niter > maxiter){
      converged = F
      con = T
      return(ModelLnLikelyhood = -Inf, converged = F)
    }
  }
}


BinomEMwrapperParallel = function(N, S, k, maxiter = 1000, maxtol = 0.0001, ntrys = 10, ncores =
parallel::detectCores()/2){
  require(foreach)
  # k is number of components/models to by fitted
  # maxiter and maxtol are thresholds for exiting the optimization
  cl = parallel::makeCluster(ncores)
  parallel::clusterExport(cl, c(lsf.str()))
  doParallel::registerDoParallel(cl)
  fit = foreach(i=1:ntrys, .packages = "Rfast", .combine = "ForeachCombineHandler") %dopar% {
      source("/Users/streeck/Desktop/EM-Project/EMcalc.R")
      return(tryCatch(BinomMultiDimEMcycle(N, S, k), error = function(w){list(ModelLnLikelyhood = -Inf)}))
    }
  parallel::stopCluster(cl)
  fit$Group = as.character(rowMaxs(fit$Expectaion))
  return(fit)
}
```

Code Snippet 1: Binomial Expectation Maximization function

7.4.3 GEN SET ENRICHMENT ANALYSIS

```
MyMultiGesa = function(sorter, featurer, cnames, outxlab = "Gene Rank by Log2 Fold Change", RankingType
= "Log Fold Change", lineplotdensity = T, niter = 10000){
  require(ggplot2)
  orderer = order(sorter, decreasing = T)
  featurer = as.matrix(featurer)
  sorter.sorted = sorter[orderer]
  featurer.sorted = as.matrix(featurer[orderer,])


  xaxis = 1:length(sorter)
  ES = matrix(data = NA, nrow = dim(featurer.sorted)[1], ncol = dim(featurer.sorted)[2])
  colnames(ES) = cnames
  t = colsums(featurer.sorted)


  for(i in 1:(dim(featurer.sorted)[2])){
    ES[featurer.sorted[,i],i] = abs(sorter.sorted[featurer.sorted[,i]])/sum(abs(sorter.sorted[featurer.sorted[,i]]))
    ES[!featurer.sorted[,i],i] = -1/(dim(featurer.sorted)[1]-t[i])
    ES[,i] = cumsum(ES[,i])
  }
  ESs = colMaxs(ES, value = T)
  iters = matrix(nrow = niter, ncol = dim(featurer.sorted)[2])
  pvals = numeric(length = dim(featurer.sorted)[2])
  names(pvals) = cnames
  for(j in 1:(dim(featurer.sorted)[2])){
    for (i in 1:niter) {
      ESi = numeric(length = dim(featurer.sorted)[1])
      fselect = sample(featurer.sorted[,j])
      ESi[fselect] = abs(sorter.sorted[fselect])/sum(abs(sorter.sorted[fselect]))
      ESi[!fselect] = -1/(dim(featurer.sorted)[1]-t[j])
      iters[i,j] = max(cumsum(ESi))
    }
    pvals[j] = max(sum(iters[,j] > ESs[j]),1)/niter
  }
  print(pvals)

  plot.data = data.frame(x = xaxis, sortstat = sorter.sorted)
  plot.data = cbind(plot.data, ES)
  plot.datamelt = reshape2::melt(plot.data, id.vars = c(1,2))

  top = ggplot(plot.datamelt, aes(x, value, color = variable)) + geom_line(size = 1.5) + theme_classic() +
ylab("Gene Set Enrichment Score") + xlab("Gene Rank") + theme(axis.title.x=element_blank(),
axis.text.x=element_blank(), panel.grid.major = element_line(color = "grey", linetype = 15)) +
    scale_color_brewer(palette = "Dark2") + theme(legend.justification = c(1.2, 1.2), legend.position = c(1, 1),
legend.title=element_blank(), axis.line.x = element_blank(), axis.ticks.x = element_blank(),
plot.margin=grid::unit(c(5.5,5.5,.5,5.5), "pt"), legend.background = element_rect(color = "black"))
  if(lineplotdensity == F){
    plot.data.lines = data.frame(x = numeric(), set = character())
    for(i in 1:(dim(featurer.sorted)[2])){
      x = xaxis[featurer.sorted[,i]]
      set = rep(cnames[i], length(x))
      plot.data.lines = rbind.data.frame(plot.data.lines, data.frame(x = x, set = set))
    }
    mid = ggplot(plot.data.lines, aes(x, set)) + geom_vline(data = plot.data.lines, aes(xintercept = x, color =
set), size = .25, alpha = .7) + theme_classic()+
      ylab("    ") + xlab("Gene Rank") + geom_blank() + xlim(min(xaxis), max(xaxis)) +
      theme(axis.ticks.y=element_blank(), axis.text.y=element_blank(), axis.title.x=element_blank(),
            axis.text.x=element_blank(), strip.background = element_blank(), strip.text.x = element_blank(),
            axis.line.x = element_blank(), axis.ticks.x = element_blank(),
            plot.margin=grid::unit(c(.5,5.5,.5,5.5), "pt"), panel.border = element_rect(size = 1, fill = NA)) +
      scale_color_brewer(palette = "Dark2", guide = F) + facet_wrap(~ set, ncol = 1)
  }
  if(lineplotdensity == T){
    if(dim(featurer)[2] > 3){
```

```
    plot.data.lines = data.frame(x = numeric(), y = numeric(), set = character(), w = numeric(), xmin =
numeric(), xmax = numeric(), ymin = numeric(), ymax = numeric())
    for(i in 1:(dim(featurer.sorted)[2])){
      d = stats::density(xaxis[featurer.sorted[,i]])
      y = d$y[d$x >=min(xaxis) & d$x <= max(xaxis)]
      x = d$x[d$x >=min(xaxis) & d$x <= max(xaxis)]
      xmin = c(min(xaxis), x[1:(length(x)-1)])
      xmax = c(x[1:(length(x)-1)], max(xaxis))
      ymin = rep(i-1, length(x))
      ymax = rep(i, length(x))
      set = rep(cnames[i], length(x))
      w = rep(max(xaxis)/length(x), length(x))
      plot.data.lines = rbind.data.frame(plot.data.lines, data.frame(x = x, y = y, set = set, w = w, xmin = xmin,
xmax = xmax, ymin = ymin, ymax = ymax))
    }
    mid = ggplot(plot.data.lines, aes(x, set, width = w)) + geom_blank() + xlim(min(xaxis), max(xaxis)) +
      geom_rect(data = plot.data.lines, aes(xmin = xmin, xmax = xmax, ymin = ymin, ymax = ymax, alpha = y,
fill = set)) +
      theme_classic()+ scale_alpha_continuous(guide = F) + scale_fill_brewer(palette = "Dark2", guide = F) +
      ylab("    ") + xlab("Gene Rank") + geom_blank() +
      theme(axis.ticks.y=element_blank(), axis.text.y=element_blank(),
            axis.title.x=element_blank(), axis.text.x=element_blank(),
            strip.background = element_blank(), strip.text.x = element_blank(),
            axis.line.x = element_blank(), axis.ticks.x = element_blank(),
            plot.margin=grid::unit(c(.5,5.5,.5,5.5), "pt"),
            panel.border = element_rect(size = 1, fill = NA),
            panel.spacing.y = grid::unit(c(0), "pt"))
  }else{
    plot.data.lines = data.frame(x = numeric(), y = numeric(), set = character(), w = numeric(), xmin =
numeric(), xmax = numeric(), ymin = numeric(), ymax = numeric())
    for(i in 1:(dim(featurer.sorted)[2])){
      d = stats::density(xaxis[featurer.sorted[,i]])
      y = d$y[d$x >=min(xaxis) & d$x <= max(xaxis)]
      x = d$x[d$x >=min(xaxis) & d$x <= max(xaxis)]
      xmin = c(min(xaxis), x[1:(length(x)-1)])
      xmax = c(x[1:(length(x)-1)], max(xaxis))
      ymin = rep(0, length(x))
      ymax = rep(1, length(x))
      set = rep(cnames[i], length(x))
      w = rep(max(xaxis)/length(x), length(x))
      plot.data.lines = rbind.data.frame(plot.data.lines, data.frame(x = x, y = y, set = set, w = w, xmin = xmin,
xmax = xmax, ymin = ymin, ymax = ymax))
    }
    mid = ggplot(plot.data.lines, aes(x, 1, width = w)) + geom_blank() + xlim(min(xaxis), max(xaxis)) +
      geom_rect(data = plot.data.lines, aes(xmin = xmin, xmax = xmax, ymin = ymin, ymax = ymax, alpha = y,
fill = set)) +
      theme_classic()+ scale_alpha_continuous(guide = F) + scale_fill_brewer(palette = "Dark2", guide = F) +
      ylab("    ") + xlab("Gene Rank") + geom_blank() +
      theme(axis.ticks.y=element_blank(), axis.text.y=element_blank(),
            axis.title.x=element_blank(), axis.text.x=element_blank(),
            strip.background = element_blank(), strip.text.x = element_blank(),
            axis.line.x = element_blank(), axis.ticks.x = element_blank(),
            plot.margin=grid::unit(c(.5,5.5,.5,5.5), "pt"),
            panel.border = element_rect(size = 1, fill = NA)) + facet_wrap(~ set, ncol = 1)
  }
 }
 bottom = ggplot(plot.data, aes(x, sortstat)) + geom_line(size = .5) + geom_area(fill = "black") +
theme_classic()+
   ylab(RankingType) + xlab(outxlab) + theme(panel.grid.major = element_line(color = "grey", linetype = 15),
plot.margin=grid::unit(c(.5,5.5,5.5,5.5), "pt"))
 gA <- ggplotGrob(top)
 gB <- ggplotGrob(mid)
 gC <- ggplotGrob(bottom)
 maxWidth = grid::unit.pmax(gA$widths[2:5], gB$widths[2:5], gC$widths[2:5])
 gA$widths[2:5] <- as.list(maxWidth)
 gB$widths[2:5] <- as.list(maxWidth)
 gC$widths[2:5] <- as.list(maxWidth)
```

```
  gridExtra::grid.arrange(gB, gA, gC, ncol=1, heights = c(1.5,7,4))
}
```

Code Snippet 2: Gene Set Enrichment Analysis function

7.4.4 QUANTILE NORMALIZATION

```
quantNorm = function(X, round = F){
  require(Rfast)
  X = as.matrix(X)
  ndim = dim(X)[2]
  OrderMat = matrix(nrow = dim(X)[1], ncol = ndim)
  for(i in 1:ndim){
    OrderMat[,i] = base::order(X[,i])
    X[,i] = X[OrderMat[,i],i]
  }
  if(round == F){
    RowMeans = Rfast::rowsums(X)/ndim
  }else{
    RowMeans = round(Rfast::rowsums(X)/ndim)
  }
  for(i in 1:ndim){
    X[OrderMat[,i],i] = RowMeans
  }
  X
}
```

Code Snippet 3: Quantile normalization function

7.4.5 GO TERM ENRICHMENT HYPERGEOMETRIC TEST

```
GOhyper = function(GAFfile, TargetSet, BackgroundSet){
  require(plyr)
  GAFfile = GAFfile[GAFfile$gene_id %in% BackgroundSet,]
  GAFfile$target = GAFfile$gene_id %in% TargetSet
  GOout = ddply(GAFfile, .(GOid), summarise, x = sum(target[!duplicated(gene_id)]), m =
length(unique(gene_id)))
  GOout$n = length(unique(BackgroundSet)) - GOout$m
  GOout$k = length(unique(TargetSet))
  GOout$logP = phyper(GOout$x-1, GOout$m, GOout$n, GOout$k, log.p = T, lower.tail = F)
  GOout$EffSize = (GOout$x/GOout$k)/(GOout$m/(GOout$n+GOout$m))
  return(GOout)
}

GOmulti = function(GAFfile, TargetSets, BackgroundSets){
  for (i in 1:length(TargetSets)) {
    GOinter = GOhyper(GAFfile, TargetSets[[i]], BackgroundSets[[i]])
    GOinter$set = names(TargetSets)[i]
    if(i == 1){
      GOout = GOinter
    }else{
      GOout = rbind(GOout, GOinter)
    }
  }
  return(GOout)
}
```

Code Snippet 4: GO term enrichment analysis using hypergeometric testing

8 APPENDIX

8.1 Bibliography

1 Jaitin, D. A. *et al.* Massively Parallel Single-Cell RNA-Seq for Marker-Free Decomposition of Tissues into Cell Types. *Science* **343,** 776-779, doi:10.1126/science.1247651 (2014).

2 Lin, S. *et al.* Comparison of the transcriptional landscapes between human and mouse tissues. *Proceedings of the National Academy of Sciences* **111,** 17224-17229, doi:10.1073/pnas.1413624111 (2014).

3 Xue, Z. *et al.* Genetic programs in human and mouse early embryos revealed by single-cell RNA sequencing. *Nature* **500,** 593-597, doi:10.1038/nature12364 (2013).

4 Latchman, D. S. Transcription factors: An overview. *The International Journal of Biochemistry & Cell Biology* **29,** 1305-1312, doi:10.1016/s1357-2725(97)00085-x (1997).

5 Seshasayee, A. S. N., Sivaraman, K. & Luscombe, N. M. 7-23 (Springer Netherlands, 2011).

6 Yusuf, D. *et al.* The Transcription Factor Encyclopedia. **13,** R24, doi:10.1186/gb-2012-13-3-r24 (2012).

7 Mondal, T., Rasmussen, M., Pandey, G. K., Isaksson, A. & Kanduri, C. Characterization of the RNA content of chromatin. *Genome Research* **20,** 899-907, doi:10.1101/gr.103473.109 (2010).

8 Luger, K., Mäder, A. W., Richmond, R. K., Sargent, D. F. & Richmond, T. J. Crystal structure of the nucleosome core particle at 2.8 Å resolution. *Nature* **389,** 251-260, doi:10.1038/38444 (1997).

9 Bednar, J. *et al.* Nucleosomes, linker DNA, and linker histone form a unique structural motif that directs the higher-order folding and compaction of chromatin. **95,** 14173-14178, doi:10.1073/pnas.95.24.14173 (1998).

10 Even-Faitelson, L., Hassan-Zadeh, V., Baghestani, Z. & Bazett-Jones, D. P. Coming to terms with chromatin structure. **125,** 95-110, doi:10.1007/s00412-015-0534-9 (2016).

11 Hu, Z. & Tee, W.-W. Enhancers and chromatin structures: regulatory hubs in gene expression and diseases. *Bioscience Reports* **37,** BSR20160183, doi:10.1042/bsr20160183 (2017).

12 Miller, T. C. & Costa, A. The architecture and function of the chromatin replication machinery. *Current Opinion in Structural Biology* **47,** 9-16, doi:10.1016/j.sbi.2017.03.011 (2017).

13 Zhang, P., Torres, K., Liu, X., Liu, C. G. & Pollock, R. E. An Overview of Chromatin-Regulating Proteins in Cells. *Curr Protein Pept Sci* **17,** 401-410 (2016).

14 Peterson, C. L. & Laniel, M.-A. Histones and histone modifications. *Current Biology* **14,** R546-R551, doi:10.1016/j.cub.2004.07.007 (2004).

15 Zhao, Y. & Garcia, B. A. Comprehensive Catalog of Currently Documented Histone Modifications. *Cold Spring Harbor Perspectives in Biology* **7,** a025064, doi:10.1101/cshperspect.a025064 (2015).

16 Barski, A. *et al.* High-Resolution Profiling of Histone Methylations in the Human Genome. *Cell* **129,** 823-837, doi:10.1016/j.cell.2007.05.009 (2007).

17 Wang, Z. *et al.* Combinatorial patterns of histone acetylations and methylations in the human genome. **40,** 897-903, doi:10.1038/ng.154 (2008).

18 Thomas, C. E., Kelleher, N. L. & Mizzen, C. A. Mass Spectrometric Characterization of Human Histone H3: A Bird's Eye View. *Journal of Proteome Research* **5,** 240-247, doi:10.1021/pr050266a (2006).

19 Loyola, A., Bonaldi, T., Roche, D., Imhof, A. & Almouzni, G. PTMs on H3 Variants before Chromatin Assembly Potentiate Their Final Epigenetic State. *Molecular Cell* **24,** 309-316, doi:https://doi.org/10.1016/j.molcel.2006.08.019 (2006).

20 Bell, O. *et al.* Localized H3K36 methylation states define histone H4K16 acetylation during transcriptional elongation in Drosophila. *Embo j* **26,** 4974-4984, doi:10.1038/sj.emboj.7601926 (2007).

21 Hallson, G. *et al.* dSet1 is the main H3K4 di- and tri-methyltransferase throughout Drosophila development. *Genetics* **190,** 91-100, doi:10.1534/genetics.111.135863 (2012).

22 Herz, H. M. *et al.* Enhancer-associated H3K4 monomethylation by Trithorax-related, the Drosophila homolog of mammalian Mll3/Mll4. *Genes Dev* **26,** 2604-2620, doi:10.1101/gad.201327.112 (2012).

23 Ketel, C. S. *et al.* Subunit contributions to histone methyltransferase activities of fly and worm polycomb group complexes. *Mol Cell Biol* **25,** 6857-6868, doi:10.1128/mcb.25.16.6857-6868.2005 (2005).

24 Mukai, M. *et al.* H3K36 Trimethylation-Mediated Epigenetic Regulation is Activated by Bam and Promotes Germ Cell Differentiation During Early Oogenesis in Drosophila. *Biol Open* **4**, 119-124, doi:10.1242/bio.201410850 (2015).

25 Coleman, R. T. & Struhl, G. Causal role for inheritance of H3K27me3 in maintaining the OFF state of a Drosophila HOX gene. *Science* **356**, eaai8236, doi:10.1126/science.aai8236 (2017).

26 Mizzen, C. A. *et al.* The TAF(II)250 subunit of TFIID has histone acetyltransferase activity. *Cell* **87**, 1261-1270, doi:10.1016/s0092-8674(00)81821-8 (1996).

27 Suganuma, T. *et al.* ATAC is a double histone acetyltransferase complex that stimulates nucleosome sliding. *Nat Struct Mol Biol* **15**, 364-372, doi:10.1038/nsmb.1397 (2008).

28 Tie, F. *et al.* CBP-mediated acetylation of histone H3 lysine 27 antagonizes Drosophila Polycomb silencing. *Development* **136**, 3131-3141, doi:10.1242/dev.037127 (2009).

29 Allfrey, V. G., Faulkner, R. & Mirsky, A. E. ACETYLATION AND METHYLATION OF HISTONES AND THEIR POSSIBLE ROLE IN THE REGULATION OF RNA SYNTHESIS. *Proc Natl Acad Sci U S A* **51**, 786-794, doi:10.1073/pnas.51.5.786 (1964).

30 Görisch, S. M., Wachsmuth, M., Tóth, K. F., Lichter, P. & Rippe, K. Histone acetylation increases chromatin accessibility. *Journal of cell science* **118**, 5825-5834, doi:10.1242/jcs.02689 (2005).

31 Ruan, K. *et al.* Histone H4 acetylation required for chromatin decompaction during DNA replication. *Scientific Reports* **5**, 12720, doi:10.1038/srep12720
https://www.nature.com/articles/srep12720#supplementary-information (2015).

32 Gong, F., Chiu, L. Y. & Miller, K. M. Acetylation Reader Proteins: Linking Acetylation Signaling to Genome Maintenance and Cancer. *PLoS genetics* **12**, e1006272, doi:10.1371/journal.pgen.1006272 (2016).

33 Marmorstein, R. & Zhou, M. M. Writers and readers of histone acetylation: structure, mechanism, and inhibition. *Cold Spring Harb Perspect Biol* **6**, a018762, doi:10.1101/cshperspect.a018762 (2014).

34 Chereji, R. V., Ramachandran, S., Bryson, T. D. & Henikoff, S. Precise genome-wide mapping of single nucleosomes and linkers in vivo. *Genome Biology* **19**, 19, doi:10.1186/s13059-018-1398-0 (2018).

35 O'Geen, H., Echipare, L. & Farnham, P. J. Using ChIP-seq technology to generate high-resolution profiles of histone modifications. *Methods in molecular biology (Clifton, N.J.)* **791**, 265-286, doi:10.1007/978-1-61779-316-5_20 (2011).

36 Wal, M. & Pugh, B. F. Genome-wide mapping of nucleosome positions in yeast using high-resolution MNase ChIP-Seq. *Methods in enzymology* **513**, 233-250, doi:10.1016/b978-0-12-391938-0.00010-0 (2012).

37 Koch, C. M. *et al.* The landscape of histone modifications across 1% of the human genome in five human cell lines. *Genome Res* **17**, 691-707, doi:10.1101/gr.5704207 (2007).

38 Schubeler, D. *et al.* The histone modification pattern of active genes revealed through genome-wide chromatin analysis of a higher eukaryote. *Genes Dev* **18**, 1263-1271, doi:10.1101/gad.1198204 (2004).

39 Schuettengruber, B., Bourbon, H. M., Di Croce, L. & Cavalli, G. Genome Regulation by Polycomb and Trithorax: 70 Years and Counting. *Cell* **171**, 34-57, doi:10.1016/j.cell.2017.08.002 (2017).

40 Muller, J. *et al.* Histone methyltransferase activity of a Drosophila Polycomb group repressor complex. *Cell* **111**, 197-208, doi:10.1016/s0092-8674(02)00976-5 (2002).

41 Czermin, B. *et al.* Drosophila enhancer of Zeste/ESC complexes have a histone H3 methyltransferase activity that marks chromosomal Polycomb sites. *Cell* **111**, 185-196, doi:10.1016/s0092-8674(02)00975-3 (2002).

42 Bowman, S. K. *et al.* H3K27 modifications define segmental regulatory domains in the Drosophila bithorax complex. *eLife* **3**, e02833, doi:10.7554/eLife.02833 (2014).

43 Brown, E. J. & Bachtrog, D. The chromatin landscape of Drosophila: comparisons between species, sexes, and chromosomes. *Genome Res* **24**, 1125-1137, doi:10.1101/gr.172155.114 (2014).

44 El-Sharnouby, S. *et al.* Regions of very low H3K27me3 partition the Drosophila genome into topological domains. *PloS one* **12**, e0172725, doi:10.1371/journal.pone.0172725 (2017).

45 Alekseyenko, A. A. *et al.* Heterochromatin-associated interactions of Drosophila HP1a with dADD1, HIPP1, and repetitive RNAs. *Genes Dev* **28**, 1445-1460, doi:10.1101/gad.241950.114 (2014).

46 Czermin, B. *et al.* Physical and functional association of SU(VAR)3-9 and HDAC1 in Drosophila. *EMBO reports* **2**, 915-919, doi:10.1093/embo-reports/kve210 (2001).

47 Azzaz, A. M. *et al.* Human heterochromatin protein 1alpha promotes nucleosome associations that drive chromatin condensation. *The Journal of biological chemistry* **289**, 6850-6861, doi:10.1074/jbc.M113.512137 (2014).

48 Bannister, A. J. *et al.* Selective recognition of methylated lysine 9 on histone H3 by the HP1 chromo domain. *Nature* **410**, 120-124, doi:10.1038/35065138 (2001).

49 Lachner, M., O'Carroll, D., Rea, S., Mechtler, K. & Jenuwein, T. Methylation of histone H3 lysine 9 creates a binding site for HP1 proteins. *Nature* **410**, 116-120, doi:10.1038/35065132 (2001).

50 Nakayama, J., Rice, J. C., Strahl, B. D., Allis, C. D. & Grewal, S. I. Role of histone H3 lysine 9 methylation in epigenetic control of heterochromatin assembly. *Science* **292**, 110-113, doi:10.1126/science.1060118 (2001).

51 Hoskins, R. A. *et al.* Sequence finishing and mapping of Drosophila melanogaster heterochromatin. *Science* **316**, 1625-1628, doi:10.1126/science.1139816 (2007).

52 Guelman, S. *et al.* Host cell factor and an uncharacterized SANT domain protein are stable components of ATAC, a novel dAda2A/dGcn5-containing histone acetyltransferase complex in Drosophila. *Mol Cell Biol* **26**, 871-882, doi:10.1128/mcb.26.3.871-882.2006 (2006).

53 Guelman, S. *et al.* The essential gene wda encodes a WD40 repeat subunit of Drosophila SAGA required for histone H3 acetylation. *Mol Cell Biol* **26**, 7178-7189, doi:10.1128/mcb.00130-06 (2006).

54 Yin, H., Sweeney, S., Raha, D., Snyder, M. & Lin, H. A high-resolution whole-genome map of key chromatin modifications in the adult Drosophila melanogaster. *PLoS genetics* **7**, e1002380, doi:10.1371/journal.pgen.1002380 (2011).

55 Kharchenko, P. V. *et al.* Comprehensive analysis of the chromatin landscape in Drosophila melanogaster. *Nature* **471**, 480-485, doi:10.1038/nature09725 (2011).

56 Yang, H. *et al.* Preferential dimethylation of histone H4 lysine 20 by Suv4-20. *The Journal of biological chemistry* **283**, 12085-12092, doi:10.1074/jbc.M707974200 (2008).

57 Karachentsev, D., Sarma, K., Reinberg, D. & Steward, R. PR-Set7-dependent methylation of histone H4 Lys 20 functions in repression of gene expression and is essential for mitosis. *Genes Dev* **19**, 431-435, doi:10.1101/gad.1263005 (2005).

58 Talasz, H., Lindner, H. H., Sarg, B. & Helliger, W. Histone H4-lysine 20 monomethylation is increased in promoter and coding regions of active genes and correlates with hyperacetylation. *The Journal of biological chemistry* **280**, 38814-38822, doi:10.1074/jbc.M505563200 (2005).

59 Vakoc, C. R., Sachdeva, M. M., Wang, H. & Blobel, G. A. Profile of histone lysine methylation across transcribed mammalian chromatin. *Mol Cell Biol* **26**, 9185-9195, doi:10.1128/mcb.01529-06 (2006).

60 Tie, F. *et al.* Trithorax monomethylates histone H3K4 and interacts directly with CBP to promote H3K27 acetylation and antagonize Polycomb silencing. *Development* **141**, 1129-1139, doi:10.1242/dev.102392 (2014).

61 Ardehali, M. B. *et al*. Drosophila Set1 is the major histone H3 lysine 4 trimethyltransferase with role in transcription. *Embo j* **30**, 2817-2828, doi:10.1038/emboj.2011.194 (2011).

62 Yuan, W. *et al.* H3K36 methylation antagonizes PRC2-mediated H3K27 methylation. *The Journal of biological chemistry* **286**, 7983-7989, doi:10.1074/jbc.M110.194027 (2011).

63 Tanaka, Y., Katagiri, Z., Kawahashi, K., Kioussis, D. & Kitajima, S. Trithorax-group protein ASH1 methylates histone H3 lysine 36. *Gene* **397**, 161-168, doi:10.1016/j.gene.2007.04.027 (2007).

64 Dorighi, K. M. & Tamkun, J. W. The trithorax group proteins Kismet and ASH1 promote H3K36 dimethylation to counteract Polycomb group repression in Drosophila. *Development* **140**, 4182-4192, doi:10.1242/dev.095786 (2013).

65 Meers, M. P. *et al.* Histone gene replacement reveals a post-transcriptional role for H3K36 in maintaining metazoan transcriptome fidelity. *eLife* **6**, doi:10.7554/eLife.23249 (2017).

66 Sneppen, K. & Ringrose, L. Theoretical analysis of Polycomb-Trithorax systems predicts that poised chromatin is bistable and not bivalent. *Nature communications* **10**, 2133, doi:10.1038/s41467-019-10130-2 (2019).

67 Jenuwein, T. & Allis, C. D. Translating the histone code. *Science* **293**, 1074-1080, doi:10.1126/science.1063127 (2001).

68 Strahl, B. D. & Allis, C. D. The language of covalent histone modifications. *Nature* **403**, 41-45, doi:10.1038/47412 (2000).

69 Filion, G. J. *et al.* Systematic protein location mapping reveals five principal chromatin types in Drosophila cells. *Cell* **143**, 212-224, doi:10.1016/j.cell.2010.09.009 (2010).

70 Moorman, C. *et al.* Hotspots of transcription factor colocalization in the genome of Drosophila melanogaster. *Proceedings of the National Academy of Sciences* **103**, 12027-12032, doi:10.1073/pnas.0605003103 (2006).

71 Karlić, R., Chung, H.-R., Lasserre, J., Vlahoviček, K. & Vingron, M. Histone modification levels are predictive for gene expression. *Proceedings of the National Academy of Sciences* **107**, 2926-2931, doi:10.1073/pnas.0909344107 (2010).

72 Henikoff, S. & Shilatifard, A. Histone modification: cause or cog? *Trends in genetics : TIG* **27**, 389-396, doi:10.1016/j.tig.2011.06.006 (2011).

73 Shimazu, T., Horinouchi, S. & Yoshida, M. Multiple histone deacetylases and the CREB-binding protein regulate pre-mRNA 3'-end processing. *The Journal of biological chemistry* **282,** 4470-4478, doi:10.1074/jbc.M609745200 (2007).

74 Levy, D. *et al.* Lysine methylation of the NF-kappaB subunit RelA by SETD6 couples activity of the histone methyltransferase GLP at chromatin to tonic repression of NF-kappaB signaling. *Nature immunology* **12,** 29-36, doi:10.1038/ni.1968 (2011).

75 Wapenaar, H. & Dekker, F. J. Histone acetyltransferases: challenges in targeting bi-substrate enzymes. *Clinical epigenetics* **8,** 59, doi:10.1186/s13148-016-0225-2 (2016).

76 Francis, N. J., Kingston, R. E. & Woodcock, C. L. Chromatin compaction by a polycomb group protein complex. *Science* **306,** 1574-1577, doi:10.1126/science.1100576 (2004).

77 Kundu, S. *et al.* Polycomb Repressive Complex 1 Generates Discrete Compacted Domains that Change during Differentiation. *Mol Cell* **65,** 432-446.e435, doi:10.1016/j.molcel.2017.01.009 (2017).

78 Gajan, A., Barnes, V. L., Liu, M., Saha, N. & Pile, L. A. The histone demethylase dKDM5/LID interacts with the SIN3 histone deacetylase complex and shares functional similarities with SIN3. *Epigenetics & chromatin* **9,** 4, doi:10.1186/s13072-016-0053-9 (2016).

79 Rozovskaia, T. *et al.* Trithorax and ASH1 interact directly and associate with the trithorax group-responsive bxd region of the Ultrabithorax promoter. *Mol Cell Biol* **19,** 6441-6447, doi:10.1128/mcb.19.9.6441 (1999).

80 Dai, J. *et al.* Probing nucleosome function: a highly versatile library of synthetic histone H3 and H4 mutants. *Cell* **134,** 1066-1078, doi:10.1016/j.cell.2008.07.019 (2008).

81 Hyland, E. M. *et al.* Insights into the role of histone H3 and histone H4 core modifiable residues in Saccharomyces cerevisiae. *Mol Cell Biol* **25,** 10060-10070, doi:10.1128/mcb.25.22.10060-10070.2005 (2005).

82 Marzluff, W. F., Gongidi, P., Woods, K. R., Jin, J. & Maltais, L. J. The human and mouse replication-dependent histone genes. *Genomics* **80,** 487-498 (2002).

83 Marzluff, W. F., Wagner, E. J. & Duronio, R. J. Metabolism and regulation of canonical histone mRNAs: life without a poly(A) tail. *Nature reviews. Genetics* **9,** 843-854, doi:10.1038/nrg2438 (2008).

84 McKay, D. J. *et al.* Interrogating the function of metazoan histones using engineered gene clusters. *Developmental cell* **32,** 373-386, doi:10.1016/j.devcel.2014.12.025 (2015).

85 Günesdogan, U., Jäckle, H. & Herzig, A. A genetic system to assess in vivo the functions of histones and histone modifications in higher eukaryotes. *EMBO reports* **11,** 772-776, doi:10.1038/embor.2010.124 (2010).

86 Lifton, R. P., Goldberg, M. L., Karp, R. W. & Hogness, D. S. The organization of the histone genes in Drosophila melanogaster: functional and evolutionary implications. *Cold Spring Harbor symposia on quantitative biology* **42 Pt 2,** 1047-1051, doi:10.1101/sqb.1978.042.01.105 (1978).

87 Penke, T. J. R., McKay, D. J., Strahl, B. D., Matera, A. G. & Duronio, R. J. Functional Redundancy of Variant and Canonical Histone H3 Lysine 9 Modification in Drosophila. *Genetics* **208,** 229-244, doi:10.1534/genetics.117.300480 (2018).

88 Hodl, M. & Basler, K. Transcription in the absence of histone H3.2 and H3K4 methylation. *Current biology : CB* **22,** 2253-2257, doi:10.1016/j.cub.2012.10.008 (2012).

89 Pengelly, A. R., Copur, O., Jackle, H., Herzig, A. & Muller, J. A histone mutant reproduces the phenotype caused by loss of histone-modifying factor Polycomb. *Science* **339,** 698-699, doi:10.1126/science.1231382 (2013).

90 Zhang, W. *et al.* Probing the Function of Metazoan Histones with a Systematic Library of H3 and H4 Mutants. *Developmental cell* **48,** 406-419.e405, doi:10.1016/j.devcel.2018.11.047 (2019).

91 Dietz, K. N. *et al.* The Drosophila Huntington's disease gene ortholog dhtt influences chromatin regulation during development. *Human molecular genetics* **24,** 330-345, doi:10.1093/hmg/ddu446 (2015).

92 Syrzycka, M. *et al.* Genetic and Molecular Analysis of Essential Genes in Centromeric Heterochromatin of the Left Arm of Chromosome 3 in Drosophila melanogaster. *G3 (Bethesda, Md.)* **9,** 1581-1595, doi:10.1534/g3.119.0003 (2019).

93 Günesdogan, U., Jäckle, H. & Herzig, A. Histone supply regulates S phase timing and cell cycle progression. *eLife* **3,** e02443, doi:10.7554/eLife.02443 (2014).

94 Grossniklaus, U. & Paro, R. Transcriptional silencing by polycomb-group proteins. *Cold Spring Harb Perspect Biol* **6,** a019331, doi:10.1101/cshperspect.a019331 (2014).

95 Struhl, G. Genes controlling segmental specification in the Drosophila thorax. *Proc Natl Acad Sci U S A* **79,** 7380-7384, doi:10.1073/pnas.79.23.7380 (1982).

96 Cao, R., Tsukada, Y. & Zhang, Y. Role of Bmi-1 and Ring1A in H2A ubiquitylation and Hox gene silencing. *Mol Cell* **20**, 845-854, doi:10.1016/j.molcel.2005.12.002 (2005).

97 Cao, R. & Zhang, Y. SUZ12 is required for both the histone methyltransferase activity and the silencing function of the EED-EZH2 complex. *Mol Cell* **15**, 57-67, doi:10.1016/j.molcel.2004.06.020 (2004).

98 Cao, R. *et al.* Role of histone H3 lysine 27 methylation in Polycomb-group silencing. *Science* **298**, 1039-1043, doi:10.1126/science.1076997 (2002).

99 Herz, H. M. *et al.* Polycomb repressive complex 2-dependent and -independent functions of Jarid2 in transcriptional regulation in Drosophila. *Mol Cell Biol* **32**, 1683-1693, doi:10.1128/mcb.06503-11 (2012).

100 Ohno, K., McCabe, D., Czermin, B., Imhof, A. & Pirrotta, V. ESC, ESCL and their roles in Polycomb Group mechanisms. *Mechanisms of development* **125**, 527-541, doi:10.1016/j.mod.2008.01.002 (2008).

101 Wen, P., Quan, Z. & Xi, R. The biological function of the WD40 repeat-containing protein p55/Caf1 in Drosophila. *Developmental dynamics : an official publication of the American Association of Anatomists* **241**, 455-464, doi:10.1002/dvdy.23730 (2012).

102 Chan, C. S., Rastelli, L. & Pirrotta, V. A Polycomb response element in the Ubx gene that determines an epigenetically inherited state of repression. *Embo j* **13**, 2553-2564 (1994).

103 Saurin, A. J., Shao, Z., Erdjument-Bromage, H., Tempst, P. & Kingston, R. E. A Drosophila Polycomb group complex includes Zeste and dTAFII proteins. *Nature* **412**, 655-660, doi:10.1038/35088096 (2001).

104 Hu, H. *et al.* CRL4B catalyzes H2AK119 monoubiquitination and coordinates with PRC2 to promote tumorigenesis. *Cancer cell* **22**, 781-795, doi:10.1016/j.ccr.2012.10.024 (2012).

105 Gutierrez, L. *et al.* The role of the histone H2A ubiquitinase Sce in Polycomb repression. *Development* **139**, 117-127, doi:10.1242/dev.074450 (2012).

106 Pengelly, A. R., Kalb, R., Finkl, K. & Müller, J. Transcriptional repression by PRC1 in the absence of H2A monoubiquitylation. *Genes Dev* **29**, 1487-1492, doi:10.1101/gad.265439.115 (2015).

107 de Napoles, M. *et al.* Polycomb group proteins Ring1A/B link ubiquitylation of histone H2A to heritable gene silencing and X inactivation. *Developmental cell* **7**, 663-676, doi:10.1016/j.devcel.2004.10.005 (2004).

108 Herz, H. M. *et al.* The H3K27me3 demethylase dUTX is a suppressor of Notch- and Rb-dependent tumors in Drosophila. *Mol Cell Biol* **30**, 2485-2497, doi:10.1128/mcb.01633-09 (2010).

109 Smith, E. R. *et al.* Drosophila UTX is a histone H3 Lys27 demethylase that colocalizes with the elongating form of RNA polymerase II. *Mol Cell Biol* **28**, 1041-1046, doi:10.1128/mcb.01504-07 (2008).

110 Lee, M. G. *et al.* Demethylation of H3K27 regulates polycomb recruitment and H2A ubiquitination. *Science* **318**, 447-450, doi:10.1126/science.1149042 (2007).

111 Mohan, M. *et al.* The COMPASS family of H3K4 methylases in Drosophila. *Mol Cell Biol* **31**, 4310-4318, doi:10.1128/mcb.06092-11 (2011).

112 Copur, Ö. & Müller, J. Histone Demethylase Activity of Utx Is Essential for Viability and Regulation of HOX Gene Expression in Drosophila. *Genetics* **208**, 633-637, doi:10.1534/genetics.117.300421 (2018).

113 Steffen, P. A. & Ringrose, L. What are memories made of? How Polycomb and Trithorax proteins mediate epigenetic memory. *Nature reviews. Molecular cell biology* **15**, 340-356, doi:10.1038/nrm3789 (2014).

114 Hansen, K. H. *et al.* A model for transmission of the H3K27me3 epigenetic mark. *Nature cell biology* **10**, 1291-1300, doi:10.1038/ncb1787 (2008).

115 Cao, R. & Zhang, Y. The functions of E(Z)/EZH2-mediated methylation of lysine 27 in histone H3. *Current opinion in genetics & development* **14**, 155-164, doi:10.1016/j.gde.2004.02.001 (2004).

116 Margueron, R. *et al.* Role of the polycomb protein EED in the propagation of repressive histone marks. *Nature* **461**, 762-767, doi:10.1038/nature08398 (2009).

117 Chiang, A., O'Connor, M. B., Paro, R., Simon, J. & Bender, W. Discrete Polycomb-binding sites in each parasegmental domain of the bithorax complex. *Development* **121**, 1681-1689 (1995).

118 Simon, J., Chiang, A., Bender, W., Shimell, M. J. & O'Connor, M. Elements of the Drosophila bithorax complex that mediate repression by Polycomb group products. *Developmental biology* **158**, 131-144, doi:10.1006/dbio.1993.1174 (1993).

119 Poux, S., Kostic, C. & Pirrotta, V. Hunchback-independent silencing of late Ubx enhancers by a Polycomb Group Response Element. *Embo j* **15**, 4713-4722 (1996).

120 Americo, J. *et al.* A complex array of DNA-binding proteins required for pairing-sensitive silencing by a polycomb group response element from the Drosophila engrailed gene. *Genetics* **160**, 1561-1571 (2002).

121 Rank, G., Prestel, M. & Paro, R. Transcription through intergenic chromosomal memory elements of the Drosophila bithorax complex correlates with an epigenetic switch. *Mol Cell Biol* **22**, 8026-8034, doi:10.1128/mcb.22.22.8026-8034.2002 (2002).

122 Cavalli, G. & Paro, R. The Drosophila Fab-7 chromosomal element conveys epigenetic inheritance during mitosis and meiosis. *Cell* **93**, 505-518, doi:10.1016/s0092-8674(00)81181-2 (1998).

123 Maurange, C. & Paro, R. A cellular memory module conveys epigenetic inheritance of hedgehog expression during Drosophila wing imaginal disc development. *Genes Dev* **16**, 2672-2683, doi:10.1101/gad.242702 (2002).

124 Okulski, H., Druck, B., Bhalerao, S. & Ringrose, L. Quantitative analysis of polycomb response elements (PREs) at identical genomic locations distinguishes contributions of PRE sequence and genomic environment. *Epigenetics & chromatin* **4**, 4, doi:10.1186/1756-8935-4-4 (2011).

125 Banerjee, U., Girard, J. R., Goins, L. M. & Spratford, C. M. Drosophila as a Genetic Model for Hematopoiesis. *Genetics* **211**, 367-417, doi:10.1534/genetics.118.300223 (2019).

126 Leitao, A. B. & Sucena, E. Drosophila sessile hemocyte clusters are true hematopoietic tissues that regulate larval blood cell differentiation. *eLife* **4**, doi:10.7554/eLife.06166 (2015).

127 Lebestky, T., Chang, T., Hartenstein, V. & Banerjee, U. Specification of Drosophila hematopoietic lineage by conserved transcription factors. *Science* **288**, 146-149, doi:10.1126/science.288.5463.146 (2000).

128 Ghosh, S., Singh, A., Mandal, S. & Mandal, L. Active hematopoietic hubs in Drosophila adults generate hemocytes and contribute to immune response. *Developmental cell* **33**, 478-488, doi:10.1016/j.devcel.2015.03.014 (2015).

129 Vlisidou, I. & Wood, W. Drosophila blood cells and their role in immune responses. *The FEBS journal* **282**, 1368-1382, doi:10.1111/febs.13235 (2015).

130 Gold, K. S. & Bruckner, K. Macrophages and cellular immunity in Drosophila melanogaster. *Seminars in immunology* **27**, 357-368, doi:10.1016/j.smim.2016.03.010 (2015).

131 Shia, A. K. *et al.* Toll-dependent antimicrobial responses in Drosophila larval fat body require Spatzle secreted by haemocytes. *Journal of cell science* **122**, 4505-4515, doi:10.1242/jcs.049155 (2009).

132 Stramer, B. *et al.* Live imaging of wound inflammation in Drosophila embryos reveals key roles for small GTPases during in vivo cell migration. *The Journal of cell biology* **168**, 567-573, doi:10.1083/jcb.200405120 (2005).

133 Kocks, C. *et al.* Eater, a transmembrane protein mediating phagocytosis of bacterial pathogens in Drosophila. *Cell* **123**, 335-346, doi:10.1016/j.cell.2005.08.034 (2005).

134 Lanot, R., Zachary, D., Holder, F. & Meister, M. Postembryonic hematopoiesis in Drosophila. *Developmental biology* **230**, 243-257, doi:10.1006/dbio.2000.0123 (2001).

135 Kurucz, E. *et al.* Definition of Drosophila hemocyte subsets by cell-type specific antigens. *Acta biologica Hungarica* **58 Suppl**, 95-111, doi:10.1556/ABiol.58.2007.Suppl.8 (2007).

136 Rizki, T. M., Rizki, R. M. & Grell, E. H. A mutant affecting the crystal cells inDrosophila melanogaster. *Wilhelm Roux's archives of developmental biology* **188**, 91-99, doi:10.1007/bf00848799 (1980).

137 Ramet, M., Lanot, R., Zachary, D. & Manfruelli, P. JNK signaling pathway is required for efficient wound healing in Drosophila. *Developmental biology* **241**, 145-156, doi:10.1006/dbio.2001.0502 (2002).

138 Neyen, C. *et al.* The Black cells phenotype is caused by a point mutation in the Drosophila pro-phenoloxidase 1 gene that triggers melanization and hematopoietic defects. *Developmental and comparative immunology* **50**, 166-174, doi:10.1016/j.dci.2014.12.011 (2015).

139 Binggeli, O., Neyen, C., Poidevin, M. & Lemaitre, B. Prophenoloxidase activation is required for survival to microbial infections in Drosophila. *PLoS pathogens* **10**, e1004067, doi:10.1371/journal.ppat.1004067 (2014).

140 Dudzic, J. P., Kondo, S., Ueda, R., Bergman, C. M. & Lemaitre, B. Drosophila innate immunity: regional and functional specialization of prophenoloxidases. *BMC biology* **13**, 81, doi:10.1186/s12915-015-0193-6 (2015).

141 Irving, P. *et al.* New insights into Drosophila larval haemocyte functions through genome-wide analysis. *Cellular microbiology* **7**, 335-350, doi:10.1111/j.1462-5822.2004.00462.x (2005).

142 Nam, H. J., Jang, I. H., Asano, T. & Lee, W. J. Involvement of pro-phenoloxidase 3 in lamellocyte-mediated spontaneous melanization in Drosophila. *Molecules and cells* **26**, 606-610 (2008).

143 Cerenius, L., Lee, B. L. & Soderhall, K. The proPO-system: pros and cons for its role in invertebrate immunity. *Trends in immunology* **29**, 263-271, doi:10.1016/j.it.2008.02.009 (2008).

144 Nappi, A. J., Vass, E., Frey, F. & Carton, Y. Superoxide anion generation in Drosophila during melanotic encapsulation of parasites. *European journal of cell biology* **68**, 450-456 (1995).

145 SHRESTHA, R. & GATEFF, E. Ultrastructure and Cytochemistry of the Cell Types in the Larval Hematopoietic Organs and Hemolymph of Drosophila Melanogaster: (drosophila/hematopoiesis/blool cells/ultrastructure/cytochemistry). *Development, Growth & Differentiation* **24**, 65-82 (1982).

146 Russo, J., Dupas, S., Frey, F., Carton, Y. & Brehelin, M. Insect immunity: early events in the encapsulation process of parasitoid (Leptopilina boulardi) eggs in resistant and susceptible strains of Drosophila. *Parasitology* **112 (Pt 1)**, 135-142, doi:10.1017/s0031182000065173 (1996).

147 Sorrentino, R. P., Carton, Y. & Govind, S. Cellular immune response to parasite infection in the Drosophila lymph gland is developmentally regulated. *Developmental biology* **243**, 65-80, doi:10.1006/dbio.2001.0542 (2002).

148 Sorrentino, R. P., Melk, J. P. & Govind, S. Genetic analysis of contributions of dorsal group and JAK-Stat92E pathway genes to larval hemocyte concentration and the egg encapsulation response in Drosophila. *Genetics* **166**, 1343-1356 (2004).

149 Kurucz, E. *et al.* Hemese, a hemocyte-specific transmembrane protein, affects the cellular immune response in Drosophila. *Proc Natl Acad Sci U S A* **100**, 2622-2627, doi:10.1073/pnas.0436940100 (2003).

150 Zettervall, C. J. *et al.* A directed screen for genes involved in Drosophila blood cell activation. *Proc Natl Acad Sci U S A* **101**, 14192-14197, doi:10.1073/pnas.0403789101 (2004).

151 Anderl, I. *et al.* Transdifferentiation and Proliferation in Two Distinct Hemocyte Lineages in Drosophila melanogaster Larvae after Wasp Infection. *PLoS pathogens* **12**, e1005746, doi:10.1371/journal.ppat.1005746 (2016).

152 Tepass, U., Fessler, L. I., Aziz, A. & Hartenstein, V. Embryonic origin of hemocytes and their relationship to cell death in Drosophila. *Development* **120**, 1829-1837 (1994).

153 Rehorn, K. P., Thelen, H., Michelson, A. M. & Reuter, R. A molecular aspect of hematopoiesis and endoderm development common to vertebrates and Drosophila. *Development* **122**, 4023-4031 (1996).

154 Schulz, R. A. & Fossett, N. Hemocyte development during Drosophila embryogenesis. *Methods in molecular medicine* **105**, 109-122 (2005).

155 Rugendorff, A., Younossi-Hartenstein, A. & Hartenstein, V. Embryonic origin and differentiation of the Drosophila heart. *Roux's archives of developmental biology : the official organ of the EDBO* **203**, 266-280, doi:10.1007/bf00360522 (1994).

156 Jung, S. H., Evans, C. J., Uemura, C. & Banerjee, U. The Drosophila lymph gland as a developmental model of hematopoiesis. *Development* **132**, 2521-2533, doi:10.1242/dev.01837 (2005).

157 Lebestky, T., Jung, S. H. & Banerjee, U. A Serrate-expressing signaling center controls Drosophila hematopoiesis. *Genes Dev* **17**, 348-353, doi:10.1101/gad.1052803 (2003).

158 Krzemien, J. *et al.* Control of blood cell homeostasis in Drosophila larvae by the posterior signalling centre. *Nature* **446**, 325-328, doi:10.1038/nature05650 (2007).

159 Sinenko, S. A., Mandal, L., Martinez-Agosto, J. A. & Banerjee, U. Dual role of wingless signaling in stem-like hematopoietic precursor maintenance in Drosophila. *Developmental cell* **16**, 756-763, doi:10.1016/j.devcel.2009.03.003 (2009).

160 Shim, J., Mukherjee, T. & Banerjee, U. Direct sensing of systemic and nutritional signals by haematopoietic progenitors in Drosophila. *Nature cell biology* **14**, 394-400, doi:10.1038/ncb2453 (2012).

161 Qiu, P., Pan, P. C. & Govind, S. A role for the Drosophila Toll/Cactus pathway in larval hematopoiesis. *Development* **125**, 1909-1920 (1998).

162 Makhijani, K., Alexander, B., Tanaka, T., Rulifson, E. & Bruckner, K. The peripheral nervous system supports blood cell homing and survival in the Drosophila larva. *Development* **138**, 5379-5391, doi:10.1242/dev.067322 (2011).

163 Holz, A., Bossinger, B., Strasser, T., Janning, W. & Klapper, R. The two origins of hemocytes in Drosophila. *Development* **130**, 4955-4962, doi:10.1242/dev.00702 (2003).

164 Yang, H., Kronhamn, J., Ekstrom, J. O., Korkut, G. G. & Hultmark, D. JAK/STAT signaling in Drosophila muscles controls the cellular immune response against parasitoid infection. *EMBO reports* **16**, 1664-1672, doi:10.15252/embr.201540277 (2015).

165 Grigorian, M., Mandal, L. & Hartenstein, V. Hematopoiesis at the onset of metamorphosis: terminal differentiation and dissociation of the Drosophila lymph gland. *Development genes and evolution* **221**, 121-131, doi:10.1007/s00427-011-0364-6 (2011).

166 Mackenzie, D. K., Bussiere, L. F. & Tinsley, M. C. Senescence of the cellular immune response in Drosophila melanogaster. *Experimental gerontology* **46**, 853-859, doi:10.1016/j.exger.2011.07.004 (2011).

167 Evans, C. J., Liu, T. & Banerjee, U. Drosophila hematopoiesis: Markers and methods for molecular genetic analysis. *Methods (San Diego, Calif.)* **68**, 242-251, doi:10.1016/j.ymeth.2014.02.038 (2014).

168 Kurucz, E. *et al.* Nimrod, a putative phagocytosis receptor with EGF repeats in Drosophila plasmatocytes. *Current biology : CB* **17**, 649-654, doi:10.1016/j.cub.2007.02.041 (2007).

169 Sears, H. C., Kennedy, C. J. & Garrity, P. A. Macrophage-mediated corpse engulfment is required for normal Drosophila CNS morphogenesis. *Development* **130**, 3557-3565, doi:10.1242/dev.00586 (2003).

170 Zhou, L., Hashimi, H., Schwartz, L. M. & Nambu, J. R. Programmed cell death in the Drosophila central nervous system midline. *Current biology : CB* **5**, 784-790, doi:10.1016/s0960-9822(95)00155-2 (1995).

171 Kiger, J. A., Jr., Natzle, J. E. & Green, M. M. Hemocytes are essential for wing maturation in Drosophila melanogaster. *Proc Natl Acad Sci U S A* **98**, 10190-10195, doi:10.1073/pnas.181338998 (2001).

172 Williams, D. W. & Truman, J. W. Cellular mechanisms of dendrite pruning in Drosophila: insights from in vivo time-lapse of remodeling dendritic arborizing sensory neurons. *Development* **132**, 3631-3642, doi:10.1242/dev.01928 (2005).

173 Nelliot, A., Bond, N. & Hoshizaki, D. K. Fat-body remodeling in Drosophila melanogaster. *Genesis (New York, N.Y. : 2000)* **44**, 396-400, doi:10.1002/dvg.20229 (2006).

174 Franc, N. C., Dimarcq, J. L., Lagueux, M., Hoffmann, J. & Ezekowitz, R. A. Croquemort, a novel Drosophila hemocyte/macrophage receptor that recognizes apoptotic cells. *Immunity* **4**, 431-443, doi:10.1016/s1074-7613(00)80410-0 (1996).

175 Manaka, J. *et al.* Draper-mediated and phosphatidylserine-independent phagocytosis of apoptotic cells by Drosophila hemocytes/macrophages. *The Journal of biological chemistry* **279**, 48466-48476, doi:10.1074/jbc.M408597200 (2004).

176 Kurant, E., Axelrod, S., Leaman, D. & Gaul, U. Six-microns-under acts upstream of Draper in the glial phagocytosis of apoptotic neurons. *Cell* **133**, 498-509, doi:10.1016/j.cell.2008.02.052 (2008).

177 Ramet, M., Manfruelli, P., Pearson, A., Mathey-Prevot, B. & Ezekowitz, R. A. Functional genomic analysis of phagocytosis and identification of a Drosophila receptor for E. coli. *Nature* **416**, 644-648, doi:10.1038/nature735 (2002).

178 Bou Aoun, R. *et al.* Analysis of thioester-containing proteins during the innate immune response of Drosophila melanogaster. *Journal of innate immunity* **3**, 52-64, doi:10.1159/000321554 (2011).

179 Lagueux, M., Perrodou, E., Levashina, E. A., Capovilla, M. & Hoffmann, J. A. Constitutive expression of a complement-like protein in toll and JAK gain-of-function mutants of Drosophila. *Proc Natl Acad Sci U S A* **97**, 11427-11432, doi:10.1073/pnas.97.21.11427 (2000).

180 Razzell, W., Evans, I. R., Martin, P. & Wood, W. Calcium flashes orchestrate the wound inflammatory response through DUOX activation and hydrogen peroxide release. *Current biology : CB* **23**, 424-429, doi:10.1016/j.cub.2013.01.058 (2013).

181 Moreira, S., Stramer, B., Evans, I., Wood, W. & Martin, P. Prioritization of competing damage and developmental signals by migrating macrophages in the Drosophila embryo. *Current biology : CB* **20**, 464-470, doi:10.1016/j.cub.2010.01.047 (2010).

182 Goto, A., Kadowaki, T. & Kitagawa, Y. Drosophila hemolectin gene is expressed in embryonic and larval hemocytes and its knock down causes bleeding defects. *Developmental biology* **264**, 582-591, doi:10.1016/j.ydbio.2003.06.001 (2003).

183 Babcock, D. T. *et al.* Circulating blood cells function as a surveillance system for damaged tissue in Drosophila larvae. *Proc Natl Acad Sci U S A* **105**, 10017-10022, doi:10.1073/pnas.0709951105 (2008).

184 Shibata, T., Hadano, J., Kawasaki, D., Dong, X. & Kawabata, S. I. Drosophila TG-A transglutaminase is secreted via an unconventional Golgi-independent mechanism involving exosomes and two types of fatty acylations. *The Journal of biological chemistry* **292**, 10723-10734, doi:10.1074/jbc.M117.779710 (2017).

185 Scherfer, C. *et al.* Isolation and characterization of hemolymph clotting factors in Drosophila melanogaster by a pullout method. *Current biology : CB* **14**, 625-629, doi:10.1016/j.cub.2004.03.030 (2004).

186 Martinek, N., Shahab, J., Saathoff, M. & Ringuette, M. Haemocyte-derived SPARC is required for collagen-IV-dependent stability of basal laminae in Drosophila embryos. *Journal of cell science* **121**, 1671-1680, doi:10.1242/jcs.021931 (2008).

187 Van De Bor, V. *et al.* Companion Blood Cells Control Ovarian Stem Cell Niche Microenvironment and Homeostasis. *Cell reports* **13**, 546-560, doi:10.1016/j.celrep.2015.09.008 (2015).

188 Kusche-Gullberg, M., Garrison, K., MacKrell, A. J., Fessler, L. I. & Fessler, J. H. Laminin A chain: expression during Drosophila development and genomic sequence. *Embo j* **11**, 4519-4527 (1992).

189 Fogerty, F. J. *et al.* Tiggrin, a novel Drosophila extracellular matrix protein that functions as a ligand for Drosophila alpha PS2 beta PS integrins. *Development* **120**, 1747-1758 (1994).

190 Olofsson, B. & Page, D. T. Condensation of the central nervous system in embryonic Drosophila is inhibited by blocking hemocyte migration or neural activity. *Developmental biology* **279**, 233-243, doi:10.1016/j.ydbio.2004.12.020 (2005).

191 Bunt, S. *et al.* Hemocyte-secreted type IV collagen enhances BMP signaling to guide renal tubule morphogenesis in Drosophila. *Developmental cell* **19**, 296-306, doi:10.1016/j.devcel.2010.07.019 (2010).

192 Johansson, K. C., Metzendorf, C. & Söderhäll, K. Microarray analysis of immune challenged Drosophila hemocytes. *Experimental cell research* **305**, 145-155 (2005).

193 Boulay, J.-L., O'Shea, J. J. & Paul, W. E. Molecular phylogeny within type I cytokines and their cognate receptors. *Immunity* **19**, 159-163 (2003).

194 Agaisse, H., Petersen, U.-M., Boutros, M., Mathey-Prevot, B. & Perrimon, N. Signaling role of hemocytes in Drosophila JAK/STAT-dependent response to septic injury. *Developmental cell* **5**, 441-450 (2003).

195 Igaki, T. *et al.* Eiger, a TNF superfamily ligand that triggers the Drosophila JNK pathway. *The EMBO journal* **21**, 3009-3018 (2002).

196 Mabery, E. M. & Schneider, D. S. The Drosophila TNF ortholog eiger is required in the fat body for a robust immune response. *Journal of innate immunity* **2**, 371-378 (2010).

197 Schneider, D. S. *et al.* Drosophila eiger mutants are sensitive to extracellular pathogens. *PLoS pathogens* **3**, e41 (2007).

198 Weber, A. N. *et al*. Binding of the Drosophila cytokine Spätzle to Toll is direct and establishes signaling. *Nature immunology* **4**, 794 (2003).

199 Mulinari, S., Häcker, U. & Castillejo-López, C. Expression and regulation of Spätzle - processing enzyme in Drosophila. *FEBS letters* **580**, 5406-5410 (2006).

200 Gangloff, M. *et al.* Structural insight into the mechanism of activation of the Toll receptor by the dimeric ligand Spätzle. *Journal of Biological Chemistry* **283**, 14629-14635 (2008).

201 Hetru, C. & Hoffmann, J. A. NF-kappaB in the immune response of Drosophila. *Cold Spring Harb Perspect Biol* **1**, a000232, doi:10.1101/cshperspect.a000232 (2009).

202 Kawai, T. & Akira, S. The role of pattern-recognition receptors in innate immunity: update on Toll-like receptors. *Nature immunology* **11**, 373-384, doi:10.1038/ni.1863 (2010).

203 Park, B. S. *et al.* The structural basis of lipopolysaccharide recognition by the TLR4–MD-2 complex. *Nature* **458**, 1191 (2009).

204 Jang, I. H. *et al.* A Spatzle-processing enzyme required for toll signaling activation in Drosophila innate immunity. *Developmental cell* **10**, 45-55, doi:10.1016/j.devcel.2005.11.013 (2006).

205 Lemaitre, B., Nicolas, E., Michaut, L., Reichhart, J. M. & Hoffmann, J. A. The dorsoventral regulatory gene cassette spatzle/Toll/cactus controls the potent antifungal response in Drosophila adults. *Cell* **86**, 973-983, doi:10.1016/s0092-8674(00)80172-5 (1996).

206 Morisato, D. & Anderson, K. V. The spätzle gene encodes a component of the extracellular signaling pathway establishing the dorsal-ventral pattern of the Drosophila embryo. *Cell* **76**, 677-688 (1994).

207 Stein, D., Roth, S., Vogelsang, E. & Nu, C. The polarity of the dorsoventral axis in the Drosophila embryo is defined by an extracellular signal. *Cell* **65**, 725-735 (1991).

208 DeLotto, Y. & DeLotto, R. Proteolytic processing of the Drosophila Spätzle protein by easter generates a dimeric NGF-like molecule with ventralising activity. *Mechanisms of development* **72**, 141-148 (1998).

209 Kang, D., Liu, G., Lundstrom, A., Gelius, E. & Steiner, H. A peptidoglycan recognition protein in innate immunity conserved from insects to humans. *Proc Natl Acad Sci U S A* **95**, 10078-10082, doi:10.1073/pnas.95.17.10078 (1998).

210 Werner, T. *et al.* A family of peptidoglycan recognition proteins in the fruit fly Drosophila melanogaster. *Proc Natl Acad Sci U S A* **97**, 13772-13777, doi:10.1073/pnas.97.25.13772 (2000).

211 Lee, W. J., Lee, J. D., Kravchenko, V. V., Ulevitch, R. J. & Brey, P. T. Purification and molecular cloning of an inducible gram-negative bacteria-binding protein from the silkworm, Bombyx mori. *Proc Natl Acad Sci U S A* **93**, 7888-7893, doi:10.1073/pnas.93.15.7888 (1996).

212 Gobert, V. *et al.* Dual activation of the Drosophila toll pathway by two pattern recognition receptors. *Science* **302**, 2126-2130, doi:10.1126/science.1085432 (2003).

213 Gottar, M. *et al.* Dual detection of fungal infections in Drosophila via recognition of glucans and sensing of virulence factors. *Cell* **127**, 1425-1437, doi:10.1016/j.cell.2006.10.046 (2006).

214 Tauszig-Delamasure, S., Bilak, H., Capovilla, M., Hoffmann, J. A. & Imler, J. L. Drosophila MyD88 is required for the response to fungal and Gram-positive bacterial infections. *Nature immunology* **3**, 91-97, doi:10.1038/ni747 (2002).

215 Lim, K. H. & Staudt, L. M. Toll-like receptor signaling. *Cold Spring Harb Perspect Biol* **5**, a011247, doi:10.1101/cshperspect.a011247 (2013).

216 Sun, H., Towb, P., Chiem, D. N., Foster, B. A. & Wasserman, S. A. Regulated assembly of the Toll signaling complex drives Drosophila dorsoventral patterning. *Embo j* **23**, 100-110, doi:10.1038/sj.emboj.7600033 (2004).

217 Belvin, M. P., Jin, Y. & Anderson, K. V. Cactus protein degradation mediates Drosophila dorsal-ventral signaling. *Genes Dev* **9**, 783-793, doi:10.1101/gad.9.7.783 (1995).

218 Fernandez, N. Q., Grosshans, J., Goltz, J. S. & Stein, D. Separable and redundant regulatory determinants in Cactus mediate its dorsal group dependent degradation. *Development* **128**, 2963-2974 (2001).

219 Bergmann, A. *et al.* A gradient of cytoplasmic Cactus degradation establishes the nuclear localization gradient of the dorsal morphogen in Drosophila. *Mechanisms of development* **60**, 109-123, doi:10.1016/s0925-4773(96)00607-7 (1996).

220 Reach, M. *et al.* A gradient of cactus protein degradation establishes dorsoventral polarity in the Drosophila embryo. *Developmental biology* **180**, 353-364, doi:10.1006/dbio.1996.0308 (1996).

221 Gottar, M. *et al.* The Drosophila immune response against Gram-negative bacteria is mediated by a peptidoglycan recognition protein. *Nature* **416**, 640-644, doi:10.1038/nature734 (2002).

222 Choe, K. M., Lee, H. & Anderson, K. V. Drosophila peptidoglycan recognition protein LC (PGRP-LC) acts as a signal-transducing innate immune receptor. *Proc Natl Acad Sci U S A* **102**, 1122-1126, doi:10.1073/pnas.0404952102 (2005).

223 Choe, K. M., Werner, T., Stoven, S., Hultmark, D. & Anderson, K. V. Requirement for a peptidoglycan recognition protein (PGRP) in Relish activation and antibacterial immune responses in Drosophila. *Science* **296**, 359-362, doi:10.1126/science.1070216 (2002).

224 Georgel, P. *et al.* Drosophila immune deficiency (IMD) is a death domain protein that activates antibacterial defense and can promote apoptosis. *Developmental cell* **1**, 503-514 (2001).

225 Lemaitre, B. *et al.* A recessive mutation, immune deficiency (imd), defines two distinct control pathways in the Drosophila host defense. *Proc Natl Acad Sci U S A* **92**, 9465-9469, doi:10.1073/pnas.92.21.9465 (1995).

226 Leulier, F., Rodriguez, A., Khush, R. S., Abrams, J. M. & Lemaitre, B. The Drosophila caspase Dredd is required to resist gram-negative bacterial infection. *EMBO reports* **1**, 353-358, doi:10.1093/embo-reports/kvd073 (2000).

227 Leulier, F., Vidal, S., Saigo, K., Ueda, R. & Lemaitre, B. Inducible expression of double-stranded RNA reveals a role for dFADD in the regulation of the antibacterial response in Drosophila adults. *Current biology : CB* **12**, 996-1000, doi:10.1016/s0960-9822(02)00873-4 (2002).

228 Boutros, M., Agaisse, H. & Perrimon, N. Sequential activation of signaling pathways during innate immune responses in Drosophila. *Developmental cell* **3**, 711-722 (2002).

229 Silverman, N. *et al.* Immune activation of NF-kappaB and JNK requires Drosophila TAK1. *The Journal of biological chemistry* **278**, 48928-48934, doi:10.1074/jbc.M304802200 (2003).

230 Kleino, A. *et al.* Inhibitor of apoptosis 2 and TAK1-binding protein are components of the Drosophila Imd pathway. *Embo j* **24**, 3423-3434, doi:10.1038/sj.emboj.7600807 (2005).

231 Stoven, S. *et al.* Caspase-mediated processing of the Drosophila NF-kappaB factor Relish. *Proc Natl Acad Sci U S A* **100**, 5991-5996, doi:10.1073/pnas.1035902100 (2003).

232 Lemaitre, B., Reichhart, J. M. & Hoffmann, J. A. Drosophila host defense: differential induction of antimicrobial peptide genes after infection by various classes of microorganisms. *Proc Natl Acad Sci U S A* **94**, 14614-14619, doi:10.1073/pnas.94.26.14614 (1997).

233 Busse, M. S., Arnold, C. P., Towb, P., Katrivesis, J. & Wasserman, S. A. A kappaB sequence code for pathway-specific innate immune responses. *Embo j* **26**, 3826-3835, doi:10.1038/sj.emboj.7601798 (2007).

234 Whitehouse, I. *et al.* Nucleosome mobilization catalysed by the yeast SWI/SNF complex. *Nature* **400**, 784-787, doi:10.1038/23506 (1999).

235 Bonnay, F. *et al.* Akirin specifies NF-κB selectivity of Drosophila innate immune response via chromatin remodeling. *Embo j* **33**, 2349-2362, doi:10.15252/embj.201488456 (2014).

236 Davis, C. A. *et al.* The Encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic acids research* **46**, D794-d801, doi:10.1093/nar/gkx1081 (2018).

237 Zheng, Q. F. *et al.* Reprogramming of histone methylation controls the differentiation of monocytes into macrophages. *The FEBS journal* **284**, 1309-1323, doi:10.1111/febs.14060 (2017).

238 Wei, G. *et al.* Global mapping of H3K4me3 and H3K27me3 reveals specificity and plasticity in lineage fate determination of differentiating CD4+ T cells. *Immunity* **30**, 155-167, doi:10.1016/j.immuni.2008.12.009 (2009).

239 Li, X. *et al.* Demethylase Kdm6a epigenetically promotes IL-6 and IFN-beta production in macrophages. *Journal of autoimmunity* **80**, 85-94, doi:10.1016/j.jaut.2017.02.007 (2017).

240 De Santa, F. *et al.* The histone H3 lysine-27 demethylase Jmjd3 links inflammation to inhibition of polycomb-mediated gene silencing. *Cell* **130**, 1083-1094, doi:10.1016/j.cell.2007.08.019 (2007).

241 Saeed, S. *et al.* Epigenetic programming of monocyte-to-macrophage differentiation and trained innate immunity. *Science* **345**, 1251086, doi:10.1126/science.1251086 (2014).

242 Qiao, Y., Kang, K., Giannopoulou, E., Fang, C. & Ivashkiv, L. B. IFN-gamma Induces Histone 3 Lysine 27 Trimethylation in a Small Subset of Promoters to Stably Silence Gene Expression in Human Macrophages. *Cell reports* **16,** 3121-3129, doi:10.1016/j.celrep.2016.08.051 (2016).

243 Kenmoku, H., Hori, A., Kuraishi, T. & Kurata, S. A novel mode of induction of the humoral innate immune response in Drosophila larvae. *Disease models & mechanisms* **10,** 271-281, doi:10.1242/dmm.027102 (2017).

244 Hedengren, M. *et al.* Relish, a central factor in the control of humoral but not cellular immunity in Drosophila. *Mol Cell* **4,** 827-837, doi:10.1016/s1097-2765(00)80392-5 (1999).

245 Tracy, C. & Krämer, H. Isolation and Infection of Drosophila Primary Hemocytes. *Bio-protocol* **7,** doi:10.21769/BioProtoc.2300 (2017).

246 Hiroyasu, A., DeWitt, D. C. & Goodman, A. G. Extraction of Hemocytes from Drosophila melanogaster Larvae for Microbial Infection and Analysis. *Journal of visualized experiments : JoVE*, doi:10.3791/57077 (2018).

247 Clark, R. I., Woodcock, K. J., Geissmann, F., Trouillet, C. & Dionne, M. S. Multiple TGF-beta superfamily signals modulate the adult Drosophila immune response. *Current biology : CB* **21,** 1672-1677, doi:10.1016/j.cub.2011.08.048 (2011).

248 Rizki, T. M. & Rizki, R. M. The direction of evolution in the Drosophila melanogaster species subgroup based on functional analysis of the crystal cells. *Journal of Experimental Zoology* **212,** 323-328, doi:10.1002/jez.1402120304 (1980).

249 Krejcova, G. *et al.* Drosophila macrophages switch to aerobic glycolysis to mount effective antibacterial defense. *eLife* **8,** doi:10.7554/eLife.50414 (2019).

250 Nairz, M. *et al.* Iron and innate antimicrobial immunity-Depriving the pathogen, defending the host. *Journal of trace elements in medicine and biology : organ of the Society for Minerals and Trace Elements (GMS)* **48,** 118-133, doi:10.1016/j.jtemb.2018.03.007 (2018).

251 Woodcock, K. J. *et al.* Macrophage-derived upd3 cytokine causes impaired glucose homeostasis and reduced lifespan in Drosophila fed a lipid-rich diet. *Immunity* **42,** 133-144, doi:10.1016/j.immuni.2014.12.023 (2015).

252 Landt, S. G. *et al.* ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res* **22,** 1813-1831, doi:10.1101/gr.136184.111 (2012).

253 Sackerson, C., Fujioka, M. & Goto, T. The even-skipped locus is contained in a 16-kb chromatin domain. *Developmental biology* **211,** 39-52, doi:10.1006/dbio.1999.9301 (1999).

254 Zeitlinger, J. *et al.* RNA polymerase stalling at developmental control genes in the Drosophila melanogaster embryo. *Nature genetics* **39,** 1512-1516, doi:10.1038/ng.2007.26 (2007).

255 Ramírez, F. *et al.* deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic acids research* **44,** W160-W165, doi:10.1093/nar/gkw257 (2016).

256 Mikhaylichenko, O. *et al.* The degree of enhancer or promoter activity is reflected by the levels and directionality of eRNA transcription. *Genes Dev* **32,** 42-57, doi:10.1101/gad.308619.117 (2018).

257 Chereji, R. V. *et al.* Genome-wide profiling of nucleosome sensitivity and chromatin accessibility in Drosophila melanogaster. *Nucleic acids research* **44,** 1036-1051, doi:10.1093/nar/gkv978 (2016).

258 Steinhauser, S., Kurzawa, N., Eils, R. & Herrmann, C. A comprehensive comparison of tools for differential ChIP-seq analysis. *Briefings in bioinformatics* **17,** 953-966, doi:10.1093/bib/bbv110 (2016).

259 Thomas, R., Thomas, S., Holloway, A. K. & Pollard, K. S. Features that define the best ChIP-seq peak calling algorithms. *Briefings in bioinformatics* **18,** 441-450, doi:10.1093/bib/bbw035 (2017).

260 Zhang, Y. *et al.* Model-based Analysis of ChIP-Seq (MACS). *Genome Biology* **9,** R137, doi:10.1186/gb-2008-9-9-r137 (2008).

261 Kharchenko, P. V., Tolstorukov, M. Y. & Park, P. J. Design and analysis of ChIP-seq experiments for DNA-binding proteins. *Nature biotechnology* **26,** 1351-1359, doi:10.1038/nbt.1508 (2008).

262 Xu, S., Grullon, S., Ge, K. & Peng, W. Spatial clustering for identification of ChIP-enriched regions (SICER) to map regions of histone methylation patterns in embryonic stem cells. *Methods in molecular biology (Clifton, N.J.)* **1150,** 97-111, doi:10.1007/978-1-4939-0512-6_5 (2014).

263 Heinig, M. *et al.* histoneHMM: Differential analysis of histone modifications with broad genomic footprints. *BMC bioinformatics* **16,** 60, doi:10.1186/s12859-015-0491-6 (2015).

264 Helmuth, J. *et al.* normR: Regime enrichment calling for ChIP-seq data. *bioRxiv*, 082263, doi:10.1101/082263 (2016).

265 Kinkley, S. *et al.* reChIP-seq reveals widespread bivalency of H3K4me3 and H3K27me3 in CD4(+) memory T cells. *Nature communications* **7,** 12514, doi:10.1038/ncomms12514 (2016).

266 Wu, C., Yang, C., Zhao, H. & Zhu, J. On the convergence of the em algorithm: A data-adaptive analysis. *arXiv preprint arXiv:1611.00519* (2016).

267 McLachlan, G. & Krishnan, T. *The EM algorithm and extensions*. Vol. 382 (John Wiley & Sons, 2007).

268 Lloret-Llinares, M. *et al.* dKDM5/LID regulates H3K4me3 dynamics at the transcription-start site (TSS) of actively transcribed developmental genes. *Nucleic acids research* **40,** 9493-9505, doi:10.1093/nar/gks773 (2012).

269 Riddle, N. C. *et al.* Plasticity in patterns of histone modifications and chromosomal proteins in Drosophila heterochromatin. *Genome Res* **21,** 147-163, doi:10.1101/gr.110098.110 (2011).

270 Smith, C. D., Shu, S., Mungall, C. J. & Karpen, G. H. The Release 5.1 annotation of Drosophila melanogaster heterochromatin. *Science* **316,** 1586-1591, doi:10.1126/science.1139815 (2007).

271 Swenson, J. M., Colmenares, S. U., Strom, A. R., Costes, S. V. & Karpen, G. H. The composition and organization of Drosophila heterochromatin are heterogeneous and dynamic. *eLife* **5,** doi:10.7554/eLife.16096 (2016).

272 He, B. *et al.* Mapping the pericentric heterochromatin by comparative genomic hybridization analysis and chromosome deletions in Drosophila melanogaster. *Genome Res* **22,** 2507-2519, doi:10.1101/gr.137406.112 (2012).

273 dos Santos, G. *et al.* FlyBase: introduction of the Drosophila melanogaster Release 6 reference genome assembly and large-scale migration of genome annotations. *Nucleic acids research* **43,** D690-697, doi:10.1093/nar/gku1099 (2015).

274 Mount, S. M. *et al.* Splicing signals in Drosophila: intron size, information content, and consensus sequences. *Nucleic acids research* **20,** 4255-4262, doi:10.1093/nar/20.16.4255 (1992).

275 Zeng, J., Kirk, B. D., Gou, Y., Wang, Q. & Ma, J. Genome-wide polycomb target gene prediction in Drosophila melanogaster. *Nucleic acids research* **40,** 5848-5863, doi:10.1093/nar/gks209 (2012).

276 Ringrose, L., Rehmsmeier, M., Dura, J. M. & Paro, R. Genome-wide prediction of Polycomb/Trithorax response elements in Drosophila melanogaster. *Developmental cell* **5,** 759-771 (2003).

277 Arnold, C. D. *et al.* Genome-Wide Quantitative Enhancer Activity Maps Identified by STARR-seq. *Science* **339,** 1074-1077, doi:10.1126/science.1232542 (2013).

278 Giner, G. & Smyth, G. K. FRY: a fast approximation to ROAST gene set test with mean aggregated set statistics. *F1000Research* **5** (2016).

279 Wu, D. *et al.* ROAST: rotation gene set tests for complex microarray experiments. *Bioinformatics* **26,** 2176-2182, doi:10.1093/bioinformatics/btq401 (2010).

280 Subramanian, A. *et al.* Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences* **102,** 15545-15550, doi:10.1073/pnas.0506580102 (2005).

281 Kwong, C. *et al.* Stability and dynamics of polycomb target sites in Drosophila development. *PLoS genetics* **4,** e1000178, doi:10.1371/journal.pgen.1000178 (2008).

282 Diaz, A., Nellore, A. & Song, J. S. CHANCE: comprehensive software for quality control and validation of ChIP-seq data. *Genome Biology* **13,** R98, doi:10.1186/gb-2012-13-10-r98 (2012).

283 Bolstad, B. M., Irizarry, R. A., Astrand, M. & Speed, T. P. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* **19,** 185-193, doi:10.1093/bioinformatics/19.2.185 (2003).

284 Amaratunga, D. & Cabrera, J. Analysis of Data From Viral DNA Microchips. *Journal of the American Statistical Association* **96,** 1161-1170, doi:10.1198/016214501753381814 (2001).

285 Efron, B. Bootstrap Methods: Another Look at the Jackknife. *Ann. Statist.* **7,** 1-26, doi:10.1214/aos/1176344552 (1979).

286 Eckardt, S., McLaughlin, K. J. & Willenbring, H. Mouse chimeras as a system to investigate development, cell and tissue function, disease mechanisms and organ regeneration. *Cell cycle (Georgetown, Tex.)* **10,** 2091-2099, doi:10.4161/cc.10.13.16360 (2011).

287 Blair, S. S. Genetic mosaic techniques for studying Drosophila development. *Development* **130,** 5065-5072, doi:10.1242/dev.00774 (2003).

288 Chou, T.-B. & Perrimon, N. Use of a yeast site-specific recombinase to produce female germline chimeras in Drosophila. *Genetics* **131,** 643-653 (1992).

289 Xu, T. & Rubin, G. M. Analysis of genetic mosaics in developing and adult Drosophila tissues. *Development* **117,** 1223-1237 (1993).

290 Birve, A. *et al.* Su(z)12, a novel Drosophila Polycomb group gene that is conserved in vertebrates and plants. *Development* **128,** 3371-3379 (2001).

291 Lee, T. & Luo, L. Mosaic analysis with a repressible cell marker (MARCM) for Drosophila neural development. *Trends in neurosciences* **24,** 251-254, doi:10.1016/s0166-2236(00)01791-4 (2001).

292 Bhargava, V., Head, S. R., Ordoukhanian, P., Mercola, M. & Subramaniam, S. Technical Variations in Low-Input RNA-seq Methodologies. *Scientific Reports* **4,** 3678, doi:10.1038/srep03678

https://www.nature.com/articles/srep03678#supplementary-information (2014).

293 Burglin, T. R. & Affolter, M. Homeodomain proteins: an update. *Chromosoma* **125**, 497-521, doi:10.1007/s00412-015-0543-8 (2016).

294 Skunca, N., Altenhoff, A. & Dessimoz, C. Quality of computationally inferred gene ontology annotations. *PLoS computational biology* **8**, e1002533, doi:10.1371/journal.pcbi.1002533 (2012).

295 Keller, A., Backes, C. & Lenhof, H. P. Computation of significance scores of unweighted Gene Set Enrichment Analyses. *BMC bioinformatics* **8**, 290, doi:10.1186/1471-2105-8-290 (2007).

296 Morin-Poulard, I., Vincent, A. & Crozatier, M. The Drosophila JAK-STAT pathway in blood cell formation and immunity. *Jak-stat* **2**, e25700, doi:10.4161/jkst.25700 (2013).

297 Tokusumi, Y., Tokusumi, T., Stoller-Conrad, J. & Schulz, R. A. Serpent, suppressor of hairless and U-shaped are crucial regulators of hedgehog niche expression and prohemocyte maintenance during Drosophila larval hematopoiesis. *Development* **137**, 3561-3568, doi:10.1242/dev.053728 (2010).

298 Nelson, B. *et al.* Activation of Imd pathway in hemocyte confers infection resistance through humoral response in Drosophila. *Biochemical and biophysical research communications* **430**, 1120-1125, doi:10.1016/j.bbrc.2012.12.027 (2013).

299 Schmid, M. R. *et al.* Control of Drosophila blood cell activation via Toll signaling in the fat body. *PloS one* **9**, e102568, doi:10.1371/journal.pone.0102568 (2014).

300 Jain, N., Moeller, J. & Vogel, V. Mechanobiology of Macrophages: How Physical Factors Coregulate Macrophage Plasticity and Phagocytosis. *Annual Review of Biomedical Engineering* **21**, 267-297, doi:10.1146/annurev-bioeng-062117-121224 (2019).

301 Das, A. *et al.* Monocyte and macrophage plasticity in tissue repair and regeneration. *The American journal of pathology* **185**, 2596-2606, doi:10.1016/j.ajpath.2015.06.001 (2015).

302 Meng, X., Khanuja, B. S. & Ip, Y. T. Toll receptor-mediated Drosophila immune response requires Dif, an NF-kappaB factor. *Genes Dev* **13**, 792-797, doi:10.1101/gad.13.7.792 (1999).

303 Dushay, M. S., Asling, B. & Hultmark, D. Origins of immunity: Relish, a compound Rel-like gene in the antibacterial defense of Drosophila. *Proc Natl Acad Sci U S A* **93**, 10343-10347, doi:10.1073/pnas.93.19.10343 (1996).

304 Fu, X. *et al.* Estimating accuracy of RNA-Seq and microarrays with proteomics. *BMC genomics* **10**, 161, doi:10.1186/1471-2164-10-161 (2009).

305 Barrett, T. *et al.* NCBI GEO: archive for functional genomics data sets—update. *Nucleic acids research* **41**, D991-D995, doi:10.1093/nar/gks1193 (2012).

306 Celniker, S. E. *et al.* Unlocking the secrets of the genome. *Nature* **459**, 927-930, doi:10.1038/459927a (2009).

307 Butler, M. J. *et al.* Discovery of genes with highly restricted expression patterns in the Drosophila wing disc using DNA oligonucleotide microarrays. *Development* **130**, 659-670, doi:10.1242/dev.00293 (2003).

308 Kirilly, D. & Xie, T. The Drosophila ovary: an active stem cell community. *Cell Research* **17**, 15-25, doi:10.1038/sj.cr.7310123 (2007).

309 Cherbas, L. *et al.* The transcriptional diversity of 25 Drosophila cell lines. *Genome Res* **21**, 301-314, doi:10.1101/gr.112961.110 (2011).

310 Wang, W. *et al.* Polycomb Group (PcG) Proteins and Human Cancers: Multifaceted Functions and Therapeutic Implications. *Medicinal research reviews* **35**, 1220-1267, doi:10.1002/med.21358 (2015).

311 Laugesen, A., Højfeldt, J. W. & Helin, K. Role of the Polycomb Repressive Complex 2 (PRC2) in Transcriptional Regulation and Cancer. *Cold Spring Harbor perspectives in medicine* **6**, doi:10.1101/cshperspect.a026575 (2016).

312 Futscher, B. W. Epigenetic changes during cell transformation. *Advances in experimental medicine and biology* **754**, 179-194, doi:10.1007/978-1-4419-9967-2_9 (2013).

313 Lee, H. *et al.* DNA copy number evolution in Drosophila cell lines. *Genome Biology* **15**, R70, doi:10.1186/gb-2014-15-8-r70 (2014).

314 LaMere, S. A. *et al.* H3K27 Methylation Dynamics during CD4 T Cell Activation: Regulation of JAK/STAT and IL12RB2 Expression by JMJD3. *Journal of immunology (Baltimore, Md. : 1950)* **199**, 3158-3175, doi:10.4049/jimmunol.1700475 (2017).

315 Diagenode, S. *H3K27me3 polyclonal antibody - Premium*, <https://www.diagenode.com/en/p/h3k27me3-polyclonal-antibody-premium-50-mg-27-ml#> (2018).

316 Kundaje, A. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317-330, doi:10.1038/nature14248 (2015).

317 Evans, K. J. *et al.* Stable Caenorhabditis elegans chromatin domains separate broadly expressed and developmentally regulated genes. *Proceedings of the National Academy of Sciences* **113,** E7020-E7029, doi:10.1073/pnas.1608162113 (2016).

318 Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature genetics* **25,** 25-29, doi:10.1038/75556 (2000).

319 Erokhin, M. *et al.* Transcriptional read-through is not sufficient to induce an epigenetic switch in the silencing activity of Polycomb response elements. *Proc Natl Acad Sci U S A* **112,** 14930-14935, doi:10.1073/pnas.1515276112 (2015).

320 Das, C. & Tyler, J. K. Histone exchange and histone modifications during transcription and aging. *Biochimica et biophysica acta* **1819,** 332-342, doi:10.1016/j.bbagrm.2011.08.001 (2013).

321 Kraushaar, D. C. *et al.* Genome-wide incorporation dynamics reveal distinct categories of turnover for the histone variant H3.3. *Genome Biol* **14,** R121, doi:10.1186/gb-2013-14-10-r121 (2013).

322 Goldberg, A. D. *et al.* Distinct factors control histone variant H3.3 localization at specific genomic regions. *Cell* **140,** 678-691, doi:10.1016/j.cell.2010.01.003 (2010).

323 Corey, L. L., Weirich, C. S., Benjamin, I. J. & Kingston, R. E. Localized recruitment of a chromatin-remodeling activity by an activator in vivo drives transcriptional elongation. *Genes Dev* **17,** 1392-1401, doi:10.1101/gad.1071803 (2003).

324 Petruk, S. *et al.* Transcription of bxd noncoding RNAs promoted by trithorax represses Ubx in cis by transcriptional interference. *Cell* **127,** 1209-1221, doi:10.1016/j.cell.2006.10.039 (2006).

325 Smith, S. T. *et al.* Modulation of heat shock gene expression by the TAC1 chromatin-modifying complex. *Nature cell biology* **6,** 162-167, doi:10.1038/ncb1088 (2004).

326 Blackledge, N. P., Rose, N. R. & Klose, R. J. Targeting Polycomb systems to regulate gene expression: modifications to a complex story. *Nature Reviews Molecular Cell Biology* **16,** 643-649, doi:10.1038/nrm4067 (2015).

327 Riising, E. M. *et al.* Gene silencing triggers polycomb repressive complex 2 recruitment to CpG islands genome wide. *Molecular cell* **55,** 347-360 (2014).

328 Hosogane, M., Funayama, R., Nishida, Y., Nagashima, T. & Nakayama, K. Ras-induced changes in H3K27me3 occur after those in transcriptional activity. *PLoS genetics* **9,** e1003698 (2013).

329 Klose, R. J., Cooper, S., Farcas, A. M., Blackledge, N. P. & Brockdorff, N. Chromatin sampling—an emerging perspective on targeting polycomb repressor proteins. *PLoS genetics* **9,** e1003717 (2013).

330 Sachs, M. *et al.* Bivalent chromatin marks developmental regulatory genes in the mouse embryonic germline in vivo. *Cell reports* **3,** 1777-1784, doi:10.1016/j.celrep.2013.04.032 (2013).

331 Pan, G. *et al.* Whole-genome analysis of histone H3 lysine 4 and lysine 27 methylation in human embryonic stem cells. *Cell stem cell* **1,** 299-312, doi:10.1016/j.stem.2007.08.003 (2007).

332 Schertel, C. *et al.* A large-scale, in vivo transcription factor screen defines bivalent chromatin as a key property of regulatory factors mediating Drosophila wing development. *Genome Res* **25,** 514-523, doi:10.1101/gr.181305.114 (2015).

333 Bonn, S. *et al.* Tissue-specific analysis of chromatin state identifies temporal signatures of enhancer activity during embryonic development. *Nature genetics* **44,** 148-156, doi:10.1038/ng.1064 (2012).

334 Doria, A. *et al.* Autoinflammation and autoimmunity: bridging the divide. *Autoimmunity reviews* **12,** 22-30, doi:10.1016/j.autrev.2012.07.018 (2012).

335 Ozen, S. What's new in autoinflammation? *Pediatric nephrology (Berlin, Germany)* **34,** 2449-2456, doi:10.1007/s00467-018-4155-4 (2019).

336 Guillou, A., Troha, K., Wang, H., Franc, N. C. & Buchon, N. The Drosophila CD36 Homologue croquemort Is Required to Maintain Immune and Gut Homeostasis during Development and Aging. *PLoS pathogens* **12,** e1005961, doi:10.1371/journal.ppat.1005961 (2016).

337 Libert, S., Chao, Y., Chu, X. & Pletcher, S. D. Trade-offs between longevity and pathogen resistance in Drosophila melanogaster are mediated by NFkappaB signaling. *Aging cell* **5,** 533-543, doi:10.1111/j.1474-9726.2006.00251.x (2006).

338 Ifrim, D. C. *et al.* Trained immunity or tolerance: opposing functional programs induced in human monocytes after engagement of various pattern recognition receptors. *Clinical and vaccine immunology : CVI* **21,** 534-545, doi:10.1128/cvi.00688-13 (2014).

339 O'Neill, A. J. Staphylococcus aureus SH1000 and 8325-4: comparative genome sequences of key laboratory strains in staphylococcal research. *Letters in applied microbiology* **51,** 358-361, doi:10.1111/j.1472-765X.2010.02885.x (2010).

340 Novick, R. Properties of a cryptic high-frequency transducing phage in Staphylococcus aureus. *Virology* **33,** 155-166, doi:10.1016/0042-6822(67)90105-5 (1967).

341 Zerbino, D. R. *et al.* Ensembl 2018. *Nucleic acids research* **46,** D754-d761, doi:10.1093/nar/gkx1098 (2018).

342 Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29,** 15-21, doi:10.1093/bioinformatics/bts635 (2013).

343 Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078-2079, doi:10.1093/bioinformatics/btp352 (2009).

344 Robinson, J. T. *et al.* Integrative genomics viewer. *Nature biotechnology* **29,** 24-26, doi:10.1038/nbt.1754 (2011).

345 Kent, W. J., Zweig, A. S., Barber, G., Hinrichs, A. S. & Karolchik, D. BigWig and BigBed: enabling browsing of large distributed datasets. *Bioinformatics* **26,** 2204-2207, doi:10.1093/bioinformatics/btq351 (2010).

346 Wang, L., Wang, S. & Li, W. RSeQC: quality control of RNA-seq experiments. *Bioinformatics* **28,** 2184-2185, doi:10.1093/bioinformatics/bts356 (2012).

347 Ewels, P., Magnusson, M., Lundin, S. & Kaller, M. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* **32,** 3047-3048, doi:10.1093/bioinformatics/btw354 (2016).

348 Liao, Y., Smyth, G. K. & Shi, W. The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote. *Nucleic acids research* **41,** e108, doi:10.1093/nar/gkt214 (2013).

349 Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15,** 550, doi:10.1186/s13059-014-0550-8 (2014).

350 McCarthy, D. J., Chen, Y. & Smyth, G. K. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic acids research* **40,** 4288-4297, doi:10.1093/nar/gks042 (2012).

351 Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26,** 139-140, doi:10.1093/bioinformatics/btp616 (2009).

352 Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic acids research* **43,** e47-e47, doi:10.1093/nar/gkv007 (2015).

353 Durinck, S., Spellman, P. T., Birney, E. & Huber, W. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nature protocols* **4,** 1184-1191, doi:10.1038/nprot.2009.97 (2009).

354 Lawrence, M. *et al.* Software for computing and annotating genomic ranges. *PLoS computational biology* **9,** e1003118, doi:10.1371/journal.pcbi.1003118 (2013).

355 Rahman, R. *et al.* Unique transposon landscapes are pervasive across Drosophila melanogaster genomes. *Nucleic acids research* **43,** 10655-10672, doi:10.1093/nar/gkv1193 (2015).

8.2 Table of Figures

8.3 Abbreviations

	Term	Full Name	Description
	ΔHisC	Df(2L)HisC	Loss of core histone complex
A			
	AMP	Antimicrobial peptide	Class of proteins and peptides with antimicrobial activities
	ANOVA	Analysis of variance	Statistical method for detecting group differences
	AttC	Attacin-C	Antimicrobial peptide against gram-negative bacteria
B			
	BDGP	Berkeley Drosophila Genome Project	Consortium that assemble the Drosophila reference genome
C			
	C. elegans	Caenorhabditis elegans	
	Cec	Cecropin	Group of antimicrobial peptide against gram-negative bacteria
	ChIP	Chromatin ImmunoPrecipitation	Method for the isolation of protein DANN complexes
	ChIP-seq	ChIP-sequencing	ChIP followed by high throughput sequencing
	CZ	Cortical zone	Part of the Drosophila lymph gland
D			
	DAMP	Danger Associated Molecular Patterns	Host derived molecules that signal danger
	DAPI	4′,6-diamidino-2-phenylindole	Fluorescent DNA stain
	Dpt	Diptericin	Antimicrobial peptide against gram-negative bacteria
	DptB	DiptericinB	Antimicrobial peptide against gram-negative bacteria
	Drs	Drosomycin	Antimicrobial peptide against gram-positive bacteria
E			
	E. coli	Escherichia coli	Gram-negative bacterium

	E(z)	Enhancer of zeste	Protein component of PRC2, H3K27 methylase
	EM	Expectation Maximization	Algorithm for fitting mixture models, here binomial mixtures
	EtOH	ethanol	
F			
	FACS	Fluorescence-activated cell sorting	Cell sorting based on cell size and fluorescent markers
	FCS	fetal calf serum	
	FLP	flippase recombinase	Recombinase protein performing FRT site-directed recombination
	FRT	flippase recognition target	DNA sequence targeted by FLP recombination
	FSC	Forward Scatter	FACS parameter of cell size
G			
	Gal4	Gal4 transcription factor	Initiates transcription downstream of uas sequences
	Gal80	Gal80 repressor	Repressor of Gal4
	GFP	Green fluorescent protein	Aequorea victoria fluorescent protein
	GO	Gene Ontology	Gene function annotation project
	GSEA	Gene set enrichment analysis	Method for testing gene set enrichment in ranked data
H			
	H3	histone H3	canonical core histone H3
	HisGU	Histone gene unit	Drosophila gene complex of core histones H1, H2A, H2B, H3 and H4
	Hml	Hemolectin	Clotting factor expressed exclusively by hemocytes
	HMM	Hidden Markov Model	Mathematical method for data classification
I			
	IGV	Integrative Genomics Viewer	Tool for high throughput sequencing visualization
	Imd	Immune deficiency	pathway for immune signaling and protein in that pathway

K			
	kb	kilobases	DNA length (1000 base pairs)
L			
	lfc	Log fold change	Log2 ratio of signals
	LPS	Lipopolysaccharides	bacterial PAMP
M			
	MACS	Model-based Analysis for ChIP-Seq	Tool for ChIP-seq peak calling
	MARCM	Mosaic analysis with a repressible cell marker	Drosophila method for marking genetic mosaics
	Mtk	Metchnikowin	Antimicrobial peptide against fungi
	MZ	Medullary zone	Part of the Drosophila lymph gland
N			
	NF-kB	nuclear factor 'kappa-light-chain-enhancer' of activated B-cells)	Transcription factor involved in immunity and development
O			
	OD600	optical density 600 nm	measurement of bacterial concentration
	OrR	Oregon-R	wild-type drosophila strain
P			
	p	p-value	Probability of observed data if the null hypothesis was correct
	PAMP	Pathogen Associated Molecular Patterns	Pathogen derived molecules that signal danger to the host
	PBS	Phosphate-buffered saline	Isotonic buffer
	PC	Principal component	Individual dimensions determined by the PCA
	PCA	Principal component analysis	Mathematical transformation for the analysis of large data sets
	PcG	Polycomb group	Group of proteins which maintain gene repression
	PGRP	Peptidoglycan recognition protein	Group of PAMP receptors
	polyA RNA	Polyadenylated RNA	Here used as mark for enrichment of mRNA
	PPO	Prophenoloxidase	Enzyme involved in melanization

	PRC2	Polycomb Repressive Complex 2	Protein complex involved in H3K27me3 placements and PcG silencing
	PRE	Polycomb response element	Genomic location of sequence based Polycomb binding
	PSC	posterior signaling center	Part of the Drosophila lymph gland
	PTM	Post-translational modification	Enzymatic change of protein amino acids
Q			
	qPCR	quantitative polymerase chain reaction	Method for detecting DNA sequence abundance
R			
	Rel	Relish	NF-kB Transcription factor involved in immunity
	RNA-PolII	RNA polymerase II	Protein enzyme complex catalyzing mRNA production
	RNA-seq	RNA-sequencing	high throughput sequencing of cDNA converted RNA
	ROAST	rotation gene set testing	Method for testing gene set enrichment in ranked data
	ROS	reactive oxygen species	group of chemically reactive chemical species containing oxygen
	RT-qPCR	Reverse Transcription Quantitative PCR	Method for detecting single target mRNA abundance
S			
	S. aureus	Staphylococcus aureus	Gram-positive bacterium
	spz	spätzle	Signaling factor for Toll
	SSC	Side Scatter	FACS parameter of inner cell complexity
	Su(z)12	Suppressor of zeste 12	Protein component of PRC2
T			
	TEP	Thioester-containing protein	Group of immune genes
	TES	Transcription end site	Genomic positions of transcriptional termination

	TPM	Tags per million	Measure of absolute transcript abundance based on RNA-seq
	TrxG	Trithorax Group	Group of proteins which maintain active gene transcription
	TSS	Transcription start site	Genomic positions of transcriptional initiation
	Tx	Triton X-100	
U			
	uas	Upstream Activation Sequence	Sequence of Gal4 binding
	Utx	Utx histone demethylase	H3K27me3 demethylase

8.4 Selbständigkeitserklärung

Hiermit erkläre ich, die Dissertation selbstständig und nur unter Verwendung der angegebenen Hilfen und Hilfsmittel angefertigt zu haben.

Ich habe mich anderwärts nicht um einen Doktorgrad beworben und besitze keinen entsprechenden Doktorgrad.

Ich erkläre, dass ich die Dissertation oder Teile davon nicht bereits bei einer anderen wissenschaftlichen Einrichtung eingereicht habe und dass sie dort weder angenommen noch abgelehnt wurde.

Ich erkläre die Kenntnisnahme der dem Verfahren zugrunde liegenden Promotionsordnung der Mathematisch-Naturwissenschaftlichen Fakultät I der Humboldt-Universität zu Berlin vom 27. Juni 2012. Weiterhin erkläre ich, dass keine Zusammenarbeit mit gewerblichen Promotionsberaterinnen/Promotionsberatern stattgefunden hat und dass die Grundsätze der Humboldt-Universität zu Berlin zur Sicherung guter wissenschaftlicher Praxis eingehalten wurden.

Robert Streeck

8.5 Acknowledgments

I want to thank Dr. Alf Herzig for being a great supervisor and mentor. Because of your continued support and motivation to try my hand on new things I was able to learn so many new things. Also thank you for the fruitful scientific discussions.

I thank Prof. Arturo Zychlinsky for his continued support in my scientific development and for passing on the constant hunger to challenge scientific findings.

I want to thank Dr. Holly Stephenson for teaching me the ways of the hemocytes and for the countless hours of discussions spent in the fly room.

I also want to thank every other member of the group Zychlinsky. You are great and I could not have wished for a better atmosphere to work in.

Dr. Ho-Ryun Chung and Dr. Alisa Fuchs I want to thank for helping me getting started in the world of Next-Generation Sequencing.

I thank Prof. Arturo Zychlinsky, Prof. Leonie Ringrose and Prof. Daniel Schubert for consenting to read and evaluate my thesis. I further thank Prof. Ana Pombo and Dr. Elena Levashina for agreeing to judge my disputation.

Last, I want to thank my family for their support. In particular, I want to thank my wife Uliana for supporting me in everything I do, and for tackling this time of long work hours with me. I also want to thank my daughter Laura for being the motivation to finally finish this thesis.